The New Encyclopædia Britannica

Volume 18

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopædia Britannica, Inc.
Robert P. Gwinn, Chairman, Board of Directors
Peter B. Norton, President
Philip W. Goetz, Editor in Chief

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

THE UNIVERSITY OF CHICAGO

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

The *Encyclopædia Britannica* is published with the editorial advice of the faculties of the University of Chicago.

Additional advice is given by committees of members drawn from the faculties of the Australian National University, the universities of British Columbia (Can.), Cambridge (Eng.), Copenhagen (Den.), Edinburgh (Scot.), Florence (Italy), Leiden (Neth.), London (Eng.), Marburg (Ger.), Montreal (Can.), Oxford (Eng.), the Ruhr (Ger.), Sussex (Eng.), Toronto (Can.), Victoria (Can.), and Waterloo (Can.); the Complutensian University of Madrid (Spain); the Max Planck Institute for Biophysical Chemistry (Ger.); the New University of Lisbon (Port.); the School of Higher Studies in Social Sciences (Fr.); Simon Fraser University (Can.); and York University (Can.).

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991
By Encyclopædia Britannica, Inc.

© 1991
By Encyclopædia Britannica, Inc.

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

No part of this work may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in U.S.A.

Library of Congress Catalog Card Number: 89-81675
International Standard Book Number: 0-85229-529-4

CONTENTS

History of Education

Education can be thought of as the transmission of the values and accumulated knowledge of a society. In this sense, it is equivalent to what social scientists term socialization or enculturation. Children—whether conceived among New Guinea tribespeople, the Renaissance Florentines, or the middle classes of Manhattan—are born without culture. Education is designed to guide them in learning a culture, molding their behaviour in the ways of adulthood, and directing them toward their eventual role in society. In the most primitive cultures, there is often little formal learning, little of what one would ordinarily call school or classes or teachers; instead, frequently, the entire environment and all activities are viewed as school and classes, and many or all adults act as teachers. As societies grow more complex, however, the quantity of knowledge to be passed on from one generation to the next becomes more than any one person can know; and hence there must evolve more selective and efficient means of cultural transmission. The outcome is formal education—the school and the specialist called the teacher.

As society becomes ever more complex and schools become ever more institutionalized, educational experience becomes less directly related to daily life, less a matter of showing and learning in the context of the workaday world, and more abstracted from practice, more a matter of distilling, telling, and learning things out of context. This concentration of learning in a formal atmosphere allows children to learn far more of their culture than they are able to do by merely observing and imitating. As society gradually attaches more and more importance to education, it also tries to formulate the overall objectives, content, organization, and strategies of education. Literature becomes laden with advice on the rearing of the younger generation. In short, there develop philosophies and theories of education.

This article deals with the evolution of the formal teaching of knowledge and skills in all parts of the world and with the various philosophies that have inspired the resulting diverse systems. A further discussion of educational theory can be found in the article PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE. The teaching profession and the functions and methods of teachers are treated in TEACHING.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 561 and 562, and the *Index.*

The article is divided into the following sections:

Education in primitive and early civilized cultures

PREHISTORIC AND PRIMITIVE CULTURES

The term education can be applied to primitive cultures only in the sense of enculturation, which is the process of cultural transmission. A primitive person, whose culture is the totality of his universe, has a relatively fixed sense of cultural continuity and timelessness. The model of life is relatively static and absolute, and it is transmitted from one generation to another with little deviation. As for prehistoric education, it can only be inferred from educational practices in surviving primitive cultures.

The purpose of primitive education is thus to guide children to becoming good members of their tribe or band. There is a marked emphasis upon training for citizenship, because primitive people are highly concerned with the growth of individuals as tribal members and the thorough comprehension of their way of life during passage from prepuberty to postpuberty.

Participatory learning in primitive societies

Because of the variety in the countless thousands of primitive cultures, it is difficult to describe any standard and uniform characteristics of prepuberty education. Nevertheless, certain things are practiced commonly within cultures. Children actually participate in the social processes of adult activities, and their participatory learning is based upon what the American anthropologist Margaret Mead has called empathy, identification, and imitation. Primitive children, before reaching puberty, learn by doing and observing basic technical practices. Their teachers are not strangers but, rather, their immediate community.

In contrast to the spontaneous and rather unregulated imitations in prepuberty education, postpuberty education in some cultures is strictly standardized and regulated. The teaching personnel may consist of fully initiated men, often unknown to the initiate though they are his relatives in other clans. The initiation may begin with the initiate being abruptly separated from his familial group and sent to a secluded camp where he joins other initiates. The purpose of this separation is to deflect the initiate's deep attachment away from his family and to establish his emotional and social anchorage in the wider web of his culture.

The initiation "curriculum" does not usually include practical subjects. Instead, it consists of a whole set of cultural values, tribal religion, myths, philosophy, history, rituals, and other knowledge. Primitive people in some cultures regard the body of knowledge constituting the initiation curriculum as most essential to their tribal membership. Within this essential curriculum, religious instruction takes the most prominent place.

EDUCATION IN THE EARLIEST CIVILIZATIONS

The Old World civilizations of Egypt, Mesopotamia, and North China. The history of civilization started in the Middle East about 3000 BC, whereas the North China civilization began about a millennium and a half later. The Mesopotamian and Egyptian civilizations flourished almost simultaneously during the first civilizational phase (3000–1500 BC). Although these civilizations differed, they shared monumental literary achievements. The need for the perpetuation of these highly developed civilizations made writing and formal education indispensable.

Egypt. Egyptian culture and education were preserved and controlled chiefly by the priests, a powerful intellectual elite in the Egyptian theocracy who also served as the political bulwarks by preventing cultural diversity. The humanities as well as such practical subjects as science, medicine, mathematics, and geometry were in the hands of the priests, who taught in formal schools. Vocational skills relating to such fields as architecture, engineering, and sculpture were generally transmitted outside the context of formal schooling.

Priestly control of Egyptian and Babylonian education

Egyptians developed two types of formal schools for privileged youth under the supervision of governmental officials and priests: one for scribes and the other for priest trainees. At the age of five, pupils entered the writing school and continued their studies in reading and writing until the age of 16 or 17. At the age of 13 or 14, the schoolboys were also given practical training in offices for which they were being prepared. Priesthood training began at the temple college, which boys entered at the age of 17, the length of training depending upon the requirements for various priestly offices. It is not clear whether or not the practical sciences constituted a part of the systematically organized curriculum of the temple college.

Rigid method and severe discipline were applied to achieve uniformity in cultural transmission, since deviation from the traditional pattern of thought was strictly prohibited. Drill and memorization were the typical methods employed. But, as noted, Egyptians also used a work-study method in the final phase of the training for scribes.

Mesopotamia. As a civilization contemporary with Egyptian civilization, Mesopotamia developed education quite similar to that of its counterpart with respect to its purpose and training. Formal education was practical and aimed to train scribes and priests. It was extended from basic reading, writing, and religion to higher learning in law, medicine, and astrology. Generally, youth of the upper classes were prepared to become scribes, who ranged from copyists to librarians and teachers. The schools for priests were said to be as numerous as temples. This in-

dicates not only the thoroughness but also the supremacy of priestly education. Very little is known about higher education, but the advancement of the priestly work sheds light upon the extensive nature of intellectual pursuit.

As in the case of Egypt, the priests in Mesopotamia dominated the intellectual and educational domain as well as the applied. The centre of intellectual activity and training was the library, which was usually housed in a temple under the supervision of influential priests. Methods of teaching and learning were memorization, oral repetition, copying of models, and individual instruction. It is believed that the exact copying of scripts was the hardest and most strenuous and served as the test of excellence in learning. The period of education was long and rigorous, and discipline was harsh.

North China. In North China, the civilization of which began with the emergence of the Shang era, complex educational practices were in effect at a very early date. In fact, every important foundation of the formation of modern Chinese character was already established, to a great extent, more than 3,000 years ago.

Moral emphases of Chinese education

Chinese ancient formal education was distinguished by its markedly secular and moral character. Its paramount purpose was to develop a sense of moral sensitivity and duty toward people and the state. Even in the early civilizational stage, harmonious human relations, rituals, and music formed the curriculum.

Formal colleges and schools probably antedate the Chou dynasty of the 1st millennium BC, at least in the imperial capitals. Local states probably had less-organized institutions, such as halls of study, village schools, and district schools. With regard to actual methods of education, ancient Chinese learned from bamboo books and obtained moral training and practice in rituals by word of mouth and example. Rigid rote learning, which typified later Chinese education, seems to have been rather condemned. Education was regarded as the process of individual development from within.

The New World civilizations of the Maya, Aztec, and Inca. The outstanding cultural achievements of the pre-Columbian civilizations are often compared with those of Old World civilizations. The ancient Mayan calendar, which surpassed Europe's Julian calendar in accuracy, was, for example, a great accomplishment demonstrating the extraordinary degree of knowledge of astronomy and mathematics possessed by the Maya. Equally impressive are the sophistication of the Inca's calendar and their highway construction, the development of the Maya's complex writing system, and the magnificent temples of the Aztec. It is unfortunate that archaeological findings and written documents hardly shed sufficient light upon education among the Maya, Aztec, and Inca. But from available documents it is evident that these pre-Columbian civilizations developed formal education for training the nobility and priests. The major purposes of education were cultural conservation, vocational training, moral and character training, and control of cultural deviation.

The Maya. Being a highly religious culture, the Maya regarded the priesthood as one of the most influential factors in the development of their society. The priest enjoyed high prestige by virtue of his extensive knowledge, literate skills, and religious and moral leadership, and high priests served as major advisers of the rulers and the nobility. To obtain a priesthood, which was usually inherited from his father or another close relative, the trainee had to receive rigorous education in the school, where priests taught history, writing, methods of divining, medicine, and the calendar system.

Priestly control of Maya and Aztec education

Character training was one of the salient features of Mayan education. The inculcation of self-restraint, cooperative work, and moderation was highly emphasized in various stages of socialization as well as on various occasions of religious festivals. In order to develop self-discipline, the future priest endured a long period of continence and abstinence, and, to develop a sense of loyalty to community, he engaged in group labour.

The Aztec. Among the Aztec, cultural preservation relied heavily upon oral transmission and rote memorization of important events, calendrical information, and religious knowledge. Priests and noble elders, who were called conservators, were in charge of education. Since one of the important responsibilities of the conservator was to censor new poems and songs, he took the greatest care in teaching poetry, particularly divine songs.

At the *calmecac,* the school for native learning where apprenticeship started at the age of 10, the history of Mexico and the content of the historical codices were systematically taught. The *calmecac* played the most vital role in ensuring oral transmission of history through oratory, poetry, and music, which were employed to make accurate memorization of events easier and to galvanize remembrance. Visual aids, such as simple graphic representations, were used to guide recitation phases, to sustain interest, and to increase comprehension of facts and dates.

The Inca. The Inca did not possess a written or recorded language as far as is known. Like the Aztec, they also depended largely on oral transmission as a means of maintaining the preservation of their culture. Inca education was divided into two distinct categories: vocational education for common Inca and highly formalized training for the nobility. As the Inca empire was a theocratic, imperial government based upon agrarian collectivism, the rulers were concerned about the vocational training of men and women in collective agriculture. Personal freedom, life, and work were subservient to the community. At birth an individual's place in the society was strictly ordained, and at five years of age every child was taken over by the government, and his socialization and vocational training were supervised by government surrogates.

Education for the nobility consisted of a four-year program that was clearly defined in terms of the curricula and rituals. In the first year the pupils learned the Quechua language, the language of the nobility. The second year was devoted to the study of religion and the third year to learning about the quipus, a complex system of knotted coloured strings or cords used for sending messages and recording historical events. In the fourth year major attention was given to the study of history, with additional instruction in sciences, geometry, geography, and astronomy. The instructors were highly respected encyclopaedic scholars known as *amautas.* After the completion of this education, the pupils were required to pass a series of rigorous examinations in order to attain full status in the life of the Inca nobility. (N.S.)

Education in classical cultures

ANCIENT INDIA

The Hindu tradition. India is the site of one of the most ancient civilizations in the world. About the 2nd millennium BC the Aryans entered the land and came into conflict with the *dāsa*s, or the non-Aryan tribes. They defeated them, spread far and wide in the country, established large-scale settlements, and founded powerful kingdoms. In the course of time, a section of the intellectuals, the Brahmans, became priests and men of learning; another group, nobles and soldiers, became Kṣatriyas; the agricultural and trading class was called Vaiśyas; and finally the *dāsa*s were absorbed as Śūdras, or domestic servants. Such was the origin of the division of the Hindus into four varnas, or "classes." By about 500 BC, the classes became hardened into castes.

Religion was the mainspring of all activities in ancient India. It was of an all-absorbing interest and embraced not only prayer and worship but philosophy, morality, law, and government as well. Religion saturated educational ideals, too, and the study of Vedic literature was indispensable to higher castes. The stages of instruction were very well defined. During the first period, the child received elementary education at home. The beginning of secondary education and formal schooling was marked by a ritual known as the *upanayana,* or thread ceremony, which was restricted to boys only and was more or less compulsory for boys of the three higher castes. The Brahman boys had this ceremony at the age of eight, the Kṣatriya boys at the age of 11, and the Vaiśya boys at the age of 12 years. The boy would leave his father's house and enter his preceptor's *āśrama,* or home, situated amid sylvan surroundings.

The Vedic tradition in Hindu education

The *ācārya* would treat him as his own child, give him free education, and not charge anything for his boarding and lodging. The pupil had to tend the sacrificial fires, do the household work of his preceptor, and look after his cattle.

The study at this stage consisted of the recitation of the Vedic mantras, or "hymns," and the auxiliary sciences—phonetics, the rules for the performance of the sacrifices, grammar, astronomy, prosody, and etymology. The character of education, however, differed according to the needs of the caste. For a child of the priestly class, there was a definite syllabus of studies. The *trayī-vidyā,* or the knowledge of the three Vedas, the most ancient of Hindu scriptures, was obligatory for him. During the whole course at school, as at college, the student had to observe brahmacharya—that is, wearing a simple dress, living on plain food, using a hard bed, and leading a celibate life.

The period of studentship normally extended to 12 years. For those who wanted to continue their studies, there was no age limit. After finishing their education at an *āśrama,* or forest school, they would join a higher centre of learning or a university presided over by a *kulapati* (a founder of a school of thought). Advanced students would also improve their knowledge by taking part in philosophical discussions at a *parisad,* or "academy." Education was not denied to women, but normally girls were instructed at home.

The method of instruction differed according to the nature of the subject. The first duty of the student was to memorize the particular Veda of his school, with special emphasis placed on correct pronunciation. In the study of such literary subjects as law, logic, rituals, and prosody, comprehension played a very important role. A third method was the use of parables, which were employed in the personal spiritual teaching relating to the Upanishads, or conclusion of the Vedas. In higher learning, such as in the teaching of dharmashastra ("righteousness science"), the most popular and useful method was catechism—the pupil asking questions and the teacher discoursing at length on the topics referred to him. Memorization, however, played the greatest role.

The introduction of Buddhist influences. By about the end of the 6th century BC, the Vedic rituals and sacrifices had gradually developed into a highly elaborate cult that profited the priests but antagonized an increasing section of the people. Education became generally confined to the Brahmans, and the *upanayana* was being gradually discarded by the non-Brahmans. The formalism and exclusiveness of the Brahmanic system was largely responsible for the rise of two new religious orders, Buddhism and Jainism. Neither of them recognized the authority of the Vedas, and both challenged the exclusive claims of the Brahmans to priesthood. They taught through the common language of the people and gave education to all, irrespective of caste, creed, or sex. Buddhism also introduced the monastic system of education. Monasteries attached to Buddhist temples served the double purpose of imparting education and of training persons for priesthood. A monastery, however, educated only those who were its members. It did not admit day scholars and thus did not cater to the needs of the entire population.

Buddhism and Jainism in Indian education

Meanwhile, significant developments were taking place in the political field that had repercussions on education. The establishment of the imperialistic Nanda dynasty in about 413 BC and then of the even stronger Mauryas some 40 years later shook the very foundations of the Vedic structure of life, culture, and polity. The Brahmans in large numbers gave up their ancient occupation of teaching in their forest retreats and took to all sorts of occupations; the Kṣatriyas also abandoned their ancient calling as warriors; and the Śūdras in their turn rose from their servile occupations. These forces produced revolutionary changes in education. Schools were established in growing towns, and even day scholars were admitted. Studies were chosen freely and not according to caste. Taxila had already acquired an international reputation in the 6th century BC as a centre of advanced studies and now improved upon it. It did not possess any college or university in the modern sense of the term, but it was a great centre of learning with a number of famous teachers, each having a school of his own.

In the 3rd century BC Buddhism received a great impetus under India's most celebrated ruler, Aśoka. After his death, Buddhism evoked resistance, and a counterreformation in Hinduism began in the country. About the 1st century AD there was also a widespread lay movement among both Buddhists and Hindus. As a result of these events, Buddhist monasteries began to undertake secular as well as religious education, and there began a large growth of popular elementary education along with secondary and higher learning.

Classical India. The 500 years from the 4th century AD to the close of the 8th, under the Guptas and Harṣa and their successors, is a remarkable period in Indian history. It was the age of the universities of Nālandā and Valabhī and of the rise of Indian sciences, mathematics, and astronomy. The university at Nālandā housed a population of several thousand teachers and students, who were maintained out of the revenues from more than 100 villages. Because of its fame, Nālandā attracted students from abroad, but the admission test was so strict that only two or three out of 10 attained admission. More than 1,500 teachers discussed over 100 different dissertations every day. These covered the Vedas, logic, grammar, Buddhist and Hindu philosophy (Sankhya, Nyaya, etc.), astronomy, and medicine. Other great centres of Buddhist learning of the post-Gupta era were Vikramaśīla, Odantapurī, and Jagaddala. The achievements in science were no less significant. Āryabhaṭa in the late 5th century was the greatest mathematician of his age. He introduced the concepts of zero and decimals. Varāhamihira of the Gupta age was a profound scholar of all the sciences and arts, from botany to astronomy and from military science to civil engineering. There was also considerable development of the medical sciences. According to contemporaries, more than eight branches of medical science, including surgery and pediatrics, were practiced by the physicians.

The university at Nālandā

These were the main developments in education prior to the Muslim invasions, beginning in the 10th century. Nearly every village had its schoolmaster, who was supported from local contributions. The Hindu schools of learning, known as *pathasala*s in western India and *tol*s in Bengal, were conducted by Brahman *ācārya*s at their residence. Each imparted instruction in an advanced branch of learning and had a student enrollment of not more than 30. Larger or smaller establishments, specially endowed by rajas and other donors for the promotion of learning, also grew in number. The usual centres of learning were either some king's capital, such as Kanauj, Dhār, Mithilā, or Ujjayinī, or a holy place, such as Vārānasi, Ayodhyā, Kāñchi, or Nasik. In addition to Buddhist viharas (monasteries), there sprang up Hindu *maṭha*s (monks' residences) and temple colleges in different parts of the country. There were also *agrahāra* villages, which were given in charity to the colonies of learned Brahmans in order to enable them to discharge their scriptural duties, including teaching. Girls were usually educated at home, and vocational education was imparted through a system of apprenticeship.

Indian influences on Asia. An account of Indian education during the ancient period would be incomplete without a discussion of the influence of Indian culture on Sri Lanka and Central and Southeast Asia. It was achieved partly through cultural or trade relations and partly through political influence. Khotān in Central Asia had a famous Buddhist vihara as early as in the 1st century AD. A number of Indian scholars lived there, and many Chinese pilgrims, instead of going to India, stayed there. Indian pandits (scholars) were also invited to China and Tibet, and many Chinese and Tibetan monks studied in Buddhist viharas in India.

The process of Indianization was at its highest in Southeast Asia. Beginning in the 2nd century AD Hindu rulers reigned in Indochina and in the numerous islands of the East Indian archipelago, from Sumatra to New Guinea, for a period of 1,500 years. These regions were peopled by primitive races, who adopted the civilization of their masters. A greater India was thus established by a general fusion of cultures. Some of the inscriptions of these countries, written in flawless Sanskrit, show the influence

Indianization of Southeast Asia

of Indian culture. There are references to Indian philosophical ideas, legends, and myths and to Indian astronomical systems and measurements. Hinduism continued to wield its influence on these lands so long as the Hindus ruled in India. This influence ceased by the 15th century AD. (S.N.M.)

ANCIENT CHINA

Ancient Chinese education served the needs of a simple agricultural society with the family as the basic social organization. Paper and the writing brush had not been invented, and the "bamboo books" then recorded to be in existence were of limited use at best. Oral instruction and teaching by example were the chief methods of education.

The molding of character was a primary aim of education. Ethical teachings stressed the importance of human relations and the family as the foundation of society. Filial piety, especially emphasizing respect for the elderly, was considered to be the most important virtue. It was the responsibility of government to provide instruction so that the talented would be able to enter government service and thus perpetuate the moral and ethical foundation of society.

Education and public service

The Chou period. *Western Chou (1111–771 BC).* This was the feudal age, when the feudal states were ruled by lords who paid homage to the king of Chou and recognized him as the "Son of Heaven."

Schools were established for the sons of the nobility in the capital city of Chou and the capital cities of the feudal states. Schools for the common people were provided within the feudal states in villages and hamlets and were attended, according to written records, by men and women after their work in the fields. There were elementary and advanced schools for both the ruling classes and the common people. Separate studies for girls were concerned chiefly with homemaking and the feminine virtues that assured the stability of the family system.

The content of education for the nobility consisted of the "six arts"—rituals, music, archery, charioteering, writing, and mathematics. They constituted what may be called the "liberal education" of the period. Mere memory work was condemned. As Confucius said of the ancient spirit of education, "learning without thought is labour lost."

Eastern Chou (770–255 BC). This was a period of social change brought about by the disintegration of the feudal order, the breakdown of traditional loyalties, the rise of cities and urban civilization, and the growth of commerce.

The instability and the perplexing problems of the times challenged scholars to propose various remedies. The absence of central control facilitated independent and creative thinking. Thus appeared one of the most creative periods in China's intellectual history, when a Hundred Schools of thought vied with one another to expound their views and proposals for attaining a happy social and political order. Some urged a return to the teachings of the sages of old, while others sought better conditions by radical change. Among the major "schools" of this age were Taoism, Confucianism, Mohism, and legalism. No one school was in the ascendancy. Each major school had its followers and disciples, among whom there was a vigorous program of instruction and intellectual discussion. Most active in the establishment of private schools were Confucius and his disciples, but the Taoists, the Mohists, and the legalists also maintained teaching institutions.

The Hundred Schools of thought

Another form of educational activity was the practice of the contending feudal states of luring to their domain a large number of scholars, partly to serve as a source of ideas for enhancing the prosperity of the state and partly to gain an aura of intellectual respectability in a land where the respect for scholars had already become an established tradition. The age of political instability and social disintegration was thus an age of free and creative intellectual activity. Conscious of their importance and responsibility, the scholars developed a tradition of self-respect and fearless criticism. It was this tradition that Confucius had in mind when he said that the educated person was not a utensil to be used, and it was this spirit that the Confucian philosopher Mencius described when he said that the great man was a man of principles whom riches and position could not corrupt, whom poverty and lowliness could not swerve, whom power and force could not bend.

The teachings of the Hundred Schools and the records of the feudal states meant a marked increase in literature and, consequently, in the materials for instruction. The classical age of China, the period of the Eastern Chou, left an intellectual and educational legacy of inestimable value. Its scholars propounded theories of government and of social and individual life that were as influential in China and East Asia as the Greek philosophers of almost contemporary age were in the Western world.

The Ch'in–Han period. *Ch'in autocracy (221–206 BC).* Of the various schools of thought that arose in China's classical age, legalism was the first to be accorded official favour. The policies of the Ch'in dynasty were based on legalist principles stressing a strong state with a centralized administration. Many of its policies were so different from past practices that they incurred the criticism of scholars, especially those who upheld the examples of the ancient sages. To stop the criticism, the ruler, who called himself the first emperor, acting upon the advice of a legalist minister, decreed a clean break with the past and a banning of books on history and of classics glorifying past rulers. Numerous books were collected and burned, and hundreds of scholars were put to death.

Though condemned for the burning of books and the persecution of scholars, the Ch'in dynasty laid the foundation for a unified empire and made it possible for the next dynasty to consolidate its power and position at home and abroad. In education, the unification efforts included a reform and simplification of the written script and the adoption of a standardized script intelligible throughout the country. First steps were taken toward uniform textbooks for the primary schools. The invention of the writing brush made of hair, as well as the making of ink, led to the replacement of the clumsy stylus and bamboo slips with writing on silk.

Scholarship under the Han (206 BC–AD 220). The Han dynasty reversed many of the policies of its short-lived predecessor. The most important change was a shift from legalism to Confucianism. The banned books were now highly regarded, and the classics became the core of education. An assiduous effort was made to recover the prohibited books and to discover books and manuscripts that scholars had concealed in secret places. Much painstaking work was done in copying and editing, and the textual and interpretative studies of the Han scholars accorded a new importance to the study of the classics. The making of paper further stimulated this revival of learning. Critical examination of old texts resulted in the practice of higher criticism long before it developed in the West.

Shift to Confucianism

There were historians, philosophers, poets, artists, and other scholars of renown in the Han dynasty. Deserving special mention is Ssu-ma Ch'ien, author of a monumental history of China from the earliest times to the 1st century BC, whose high level of scholarship earned him the title "Chinese Father of History." An illustrious woman of letters, Pan Chao, was named poet laureate. A bibliographer collected and edited ancient texts and designated them as classics. The first dictionary of the Chinese language was written. Since the discovery and interpretation of ancient texts had largely been the work of Confucian scholars, Chinese scholarship from now on became increasingly identified with Confucianism. Most of the Han rulers gave official sanction to Confucianism as a basis of conducting government and state affairs. There was, however, no action to exclude other schools of thought.

There were a variety of schools on the national and local levels. Increasing activity in private education continued, and much of the study of the classics and enriched literature was done in private schools. Of considerable influence in the country and abroad was a national university with an enrollment that soared to 30,000. The classics now became the core of the curriculum, but music, rituals, and archery were still included. The tradition of all-round education in the six arts had not vanished.

Introduction of Buddhism. The Han dynasty was a period of territorial expansion and growth in trade and cultural relations. Buddhism was introduced at this time.

Buddhist and Indian influences in China

Early information about Buddhism was probably brought into China by traders, envoys, and monks. By the 1st century AD an emperor became personally interested and sent a mission to India to seek more knowledge and bring back Buddhist literature. Thereafter, Indian missionaries as well as Chinese scholars translated Buddhist scriptures and other writings into Chinese.

Indian missionaries not only preached a new faith but also brought in new cultural influences. Indian mathematics and astronomical ideas enriched Chinese knowledge in these fields. Chinese medicine also benefited. Architectural and art forms reflected Buddhist and Indian influence. Hindu chants became a part of Chinese music.

For a couple of centuries after its introduction, however, Buddhism showed no signs of popular appeal. Han scholarship was engrossed in the study of ancient classics and was dominated by Confucian scholars who had scant interest in Buddhist teachings that were unconcerned with the practical issues of moral and political life. Moreover, the Buddhist view of evil and the Buddhist espousal of celibacy and escape from earthly existence were alien to China's traditions. Taoist scholars, finding in Buddhism much that seemed not too remote from their own spiritual message, were more inclined to study the new philosophy. Some of them aided in the translation of Buddhist texts, but they were not in the centre of the Han stage.

The fall of the Han dynasty was followed by a few hundred years of division, strife, and foreign invasions. China was not united again until the end of the 6th century. It was during this period that Buddhism gained a foothold in China. The literary efforts of Chinese monks produced a Chinese Buddhist literature, and this marked the beginning of a process that transformed an alien importation into a Chinese religion and system of thought.

(T.H.C.)

ANCIENT HEBREWS

Like all preindustrial societies, ancient Israel first experienced a type of education that was essentially familial; that is to say, the mother taught the very young and the girls, while the father assumed the responsibility of providing moral, religious, and handcraft instruction for the growing sons. This characteristic remained in Jewish education, for the relation of teacher to pupil was always expressed in terms of parenthood and filiation. Education, furthermore, was rigid and exacting; the Hebrew word *musar* signifies at the same time education and corporal punishment.

The education of scribes

Once they were established in Palestine—at the crossroads of the great literate civilizations of the Middle East, in the beginning of the 1st millennium BC—the Jewish people learned to develop a different type of education—one that involved training a specialized, professional class of scribes in a then rather esoteric art called writing, borrowed from the Phoenicians. Writing was at first practical: the scribe wrote letters and drew up contracts, kept accounts, maintained records, and prepared orders. Because he could receive written orders, he eventually became entrusted with their execution; hence the importance of scribes in the royal administration, well-attested since the times of David and Solomon. The training given these scribes, moreover, included training of character and instilling the high ideal of wisdom, as would befit the servants of the king.

Writing found another avenue of application in Israel—in religion. And the scribe again was the agent of education. He was the man who copied the sacred Law faithfully and established the canonical text. He was the one who read the Law to himself and to the people, taught it, and translated it when Hebrew ceased to be the vernacular or "living language" (into Greek in Alexandria, into Aramaic in Palestine); he explained it, commented on it, and studied its application in particular cases. After the downfall of Israel in 722–721 BC and Judah in 586 BC and their subjection to foreign rule, Jewish education became characterized more and more by this religious orientation. The synagogue in which the community assembled became not merely a house of prayer but also a school, with a "house of the book" (*bet ha-sefer*) and a "house of instruction" (*bet ha-midrash*) corresponding roughly to elementary and secondary or advanced levels of education. Girls, however, continued to be taught at home.

The role of writing in this Oriental world should not be exaggerated, of course; oral instruction still held first place by far. Although a pupil might learn to read aloud, or rather to intone his text, his main effort was to learn by heart fragment after fragment of the sacred Law. Alongside this written Law, however, there developed interpretations or exegeses of it, which at first were merely oral but which progressively were reduced to writing—first in the form of memoranda or aide-mémoire inscribed on tablets or notebooks, then in actual books. The diffusion of this religious literature called for an expansion of programs of instruction, evolving into diverse stages: elementary, intermediate, and advanced, the latter in several centres in Palestine, later in Babylonia. This religiously based education was to become one of the most important factors enabling Judaism to survive the national catastrophes of AD 70 and 135, involving the capture and subsequent destruction of Jerusalem. In their dispersion, the Jews clung to Hebrew, their only language for worship, for the study of the Law, for tradition, and consequently for instruction. From this evolved the respect with which the teacher was and is surrounded in Jewish communities.

ANCIENT GREEKS

Origins. The history of the Hellenic language and therewith of the Hellenic people goes back to the Mycenaean civilization of about 1400–1100 BC, which itself was the heir of the pre-Hellenic civilization of Minoan Crete. The Mycenaean civilization consisted of little monarchies of an Oriental type, with an administration operated by a bureaucracy, and it seems to have operated an educational system designed for the training of scribes, similar to those of the ancient civilizations of the Middle East. But continuity did not exist between this education and that which was to develop after a period of obscurity known as the Greek Dark Age, dating approximately from the 11th to the 8th century BC. When the Greek world reappeared in history, it was an entirely different society, one headed by a military aristocracy as idealized in Homer's *Iliad* and *Odyssey*.

Education in the Greek Archaic period

During this period, sons of the nobility received their education at the court of the prince in the setting of a guild companionship of warriors: the young nobleman was educated through the counsel and example of an older man to whom he had been entrusted or had entrusted himself, a senior admired and loved. It was in this atmosphere of virile camaraderie that there developed the characteristic ideal of Greek love that was enduringly to mark Hellenic civilization and to deeply influence its conception of education itself—for example, in the relation of master to pupil. Yet these warriors of the Archaic period were not coarse barbarians; by this time the Homerids (reciters of Homer) and the rhapsodists (singers-reciters and sometimes creative poets) were taking the great epics of Homer and Hesiod throughout the far-flung Greek settlements of the Mediterranean, and a new, cultivated civilization was already emerging. Dance, poetry, and instrumental music were well developed and provided an essential element in the educational formation of the dominant elites. In addition, the idea of *aretē* was becoming central to Greek life. The epics of Hesiod and Homer glorified physical and military prowess and promoted the ideal of the cultivated patriot-warrior who displayed this cardinal virtue of *aretē*, a concept difficult to translate but embodying the virtues of military skill, moral excellence, and educational cultivation. It was an ethic of honour, which made virtues of pride and of jealousy as the inspiration of great deeds and which accepted it as natural that one would be the object of jealousy or of enmity. Reverence for Homer, which until the end of antiquity (and in Byzantium even later) was to constitute the basis of Greek culture and therewith of Greek education, would maintain from generation to generation this "agonistic" ideal: the cult of the hero, of the champion, of high performance, which found an outlet outside the sphere of battles in games or contests (*agōnes*), particularly in the realm of athletics, the most celebrated being the Olympic Games, dating traditionally from 776 BC.

Profound changes were introduced into Greek education as a result of the political transformations involved in the maturing of the city-state. There developed a collective ideal of devotion to the community: the city-state (*polis*) was everything to its citizens; the city made its citizens what they were—mankind. This subordination of the individual exploit to collective discipline was reinforced by the strategic military revolution that saw the triumph of heavy infantry, the hoplites, foot soldiers heavily armed and in tight formation.

Sparta. It is in Sparta, the most flourishing city of the 8th and 7th centuries BC, that one sees to best advantage the richness and complexity of this archaic culture. Education was carried to a high level of artistic refinement, as evidenced by the events organized within the framework of the city's religious festivals. The young men and women engaged in processions, dances, and competitions in instrumental music and song. Physical education had a like part, equally for both sexes, given status by national or international contests (the Spartans regularly took more than half of the first places at the Olympic Games); but military and civic education dominated, as it was expected that the citizen-soldier be ready to fight and, if necessary, to die, for his country.

This last aspect became not merely dominant but exclusive from the time (about 550 BC) when a conservative reaction triumphed at Sparta, bringing to power a militarist and aristocratic regime. Arts and sports gave way completely to an education appropriate to men of a warrior caste. The education of girls was subordinated to their future function as mothers; a strict eugenic regime pitilessly eliminated sickly and deformed children. Up to the age of seven, children were brought up by the women, already in an atmosphere of severity and harshness. Education, properly speaking, *agōgē,* lasted from age seven to 20 and was entirely in the hands of the state.

Spartan education for a warrior caste

The male youth of Sparta were enrolled into formations corresponding to successive age classes, divided into smaller units under the authority of comrades of their own age or of young officers. It was a collective education, which progressively removed them from the family and subjected them to garrison life. Everything was organized with a view to preparation for military service: lightly clothed, bedded on the bare ground, the child was poorly fed, told to steal to supplement his rations, and subjected to rigorous discipline. His virility and combativeness were developed by hardening him to blows—thus the role of ritual brawls between groups of boys and of the institution of the *krypteia,* a nocturnal expedition designed both to terrify the lower classes of slaves (helots) and to train the future fighter in ambushes and the ruses of warfare. He was also, of course, directly apprenticed to the military craft, using arms and maneuvering in close formation. This puritanical education, proceeding in a climate of austerity, had as its sole norm the interests of the state, erected into a supreme category; the Spartan was trained under a strict discipline to obey blindly the orders of his superiors. Curiously, the child was at the same time trained to dissimulation, to lying, to theft—all virtues when directed toward the foreigner, toward whom distrust and Machiavellianism were encouraged.

This implacably logical education enabled Sparta to remain for long the most powerful city, militarily and diplomatically, of the entire Greek world and to triumph over its rival Athens after the long struggle of the Peloponnesian War (431–404 BC); but it did not prevent Sparta's decadence. Not that Sparta ever relaxed its tension: on the contrary, in the course of centuries, the rigour and ferocity were accentuated even as such behaviour became more and more anachronistic and without real use. Rites of initiation were transformed into barbarous tests of endurance, the boys undergoing flagellation and competing in enduring it, sometimes to the very death, under the eyes of tourists attracted by the sadistic spectacle. This occurred in times of complete peace when, under the Roman Empire, Sparta was nothing but a little provincial city with neither independence nor army.

Athens. Beginning at a date difficult to fix precisely (at the end of the 7th or during the 6th century), Athens, in contrast to Sparta, became the first to renounce education oriented toward the future duties of the soldier. The Athenian citizen, of course, was always obliged, when necessary and capable, to fight for the fatherland, but the civil aspect of life and culture was predominant: armed combat was only a sport. The evolution of Athenian education reflected that of the city itself, which was moving toward increasing democratization—though it should be noted that the slave and the resident alien always remained excluded from the body politic. The Athenian democracy, even in its most complete form, attained in the 4th century BC, was to remain always the way of life of a minority—about one-tenth, it is estimated, of the total population. Athenian culture continued to be oriented toward the noble life, that of the Homeric knight, minus the warrior aspect, and this orientation determined the practice of elegant sports. Some of these, such as horsemanship and hunting, always remained more or less the privilege of an aristocratic and wealthy elite; the various branches of athletics, however, originally reserved for the sons of the great families, became more and more widely practiced.

Athenian education for a democratic minority

Education of youth. Schools had begun to appear in those early centuries, probably on eastern Mediterranean models, run by private teachers. The earliest references are, however, more recent. Herodotus mentions schools dating from 496 BC and Pausanias from 491 BC. The term used is *didaskaleion* ("a place for instruction"), while the generic term *scholē,* meaning leisure—a reference to schooling being the preserve of the wealthier sector—was also coming into use. There was no single institution; rather, each activity was carried out in a separate place. The young boy of privileged rank would be taken by a kind of chaperon, the *paidagōgos,* who was generally a respected slave within the parents' household. The elements of literacy were taught by the writing master, known as a *grammatistes,* the child learning his letters and numbers by scratching them on a wax-coated wooden tablet with a stylus. More advanced formal literacy, chiefly in a study of the poets, playwrights, and historians, was given by the *grammatikos,* although this was restricted to the genuinely leisured. Supremely important was instruction in the mythopoeic legends of Hesiod and Homer, given by the lyre-playing *kitharistes.* In addition, all boys had to be instructed in physical and military activities in the wrestling school, known as the palaestra, itself part of the more comprehensive institution of the gymnasium.

The moral aspect of education was not neglected. The Athenian ideal was that of the *kalos k'agathos,* the "wise and good" man. The teachers were as much preoccupied with overseeing the child's good conduct and the formation of his character as with directing his progress in the various subjects taught him. Poetry served to transmit all the traditional wisdom, which combined two currents: the ethic of the citizen expressed in the moralizing elegies of the 6th-century lawmaker Solon and the old Homeric ideal of the value of competition and heroic exploit. But this ideal equilibrium between the education of the body and that of the mind was interrupted before long as a result on the one hand of the development of professional sports and the exigencies of its specialization and on the other by the development of the strictly intellectual disciplines, which had made great progress since the time of the first philosophers of the 5th century BC.

Higher education. A system of higher education open to all—to all, at any rate, who had the leisure and necessary money—emerged with the appearance of the Sophists, mostly foreign teachers who were contemporaries and adversaries of Socrates (*c.* 470–399 BC). Until then, the higher forms of culture had retained an esoteric character, being transmitted by the master to a few chosen disciples, as in the first schools of medicine at Cnidus and at Cos, or within the framework of a religious confraternity involving initiate status. The Sophists proposed to meet a new need that was generally felt in Greek society, particularly in the most active cities such as Athens, where political life had been intensively developed. Henceforth, participation in public affairs became the supreme occupation engaging the ambition of Greek man; it was no longer in athletics and elegant leisure activities that his valour, his desire to

Sophistic education

assert himself and to triumph, would find expression but rather in political action.

The Sophists, who were professional educators, introduced a form of higher education whose commercial success attested to and was promoted by its social utility and practical efficacy. They inaugurated the literary genre of the public lecture, which was to experience a long popularity. It was a teaching process that was oriented in an entirely realistic direction, education for political participation. The Sophists pretended neither to transmit nor to seek for the truth concerning man or existence; they offered simply an art of success in political life, which meant, above all, being able on every occasion to make one's point of view prevail. Two principal disciplines constituted the program: the art of logical argument, or dialectic, and the art of persuasive speaking, or rhetoric—the two most flourishing humanistic sciences of antiquity. These disciplines the Sophists founded by distilling from experience their general principles and logical structures, thus making possible their transmission on a theoretical basis from master to pupil.

To the pedagogy of the Sophists there was opposed the activity of Socrates, who, as inheritor of the earlier aristocratic tradition, was alarmed by this radical utilitarianism. He doubted that virtue could be taught, especially for money, a degrading substance. An heir also of the old sages of former times, Socrates held that the supreme ideal of man and hence of education was not the spirit of efficiency and power but the distinterested search for the absolute, for virtue—in short, for knowledge and understanding.

It was only at the beginning of the 4th century BC, however, that the principal types of classical Greek higher education became organized on definitive lines. This was the result of the joint and rival efforts of the two great educators, the philosopher Plato (*c.* 428–348/347), who opened his school, the Academy, probably in 387, and the orator Isocrates (436–338), who founded his school in about 390.

Platonic education

Plato was descended from a long line of aristocrats and became the most distinguished of Socrates' students. The indictment and execution of Socrates by what Plato considered an ignorant society turned him away from Athens and public life. After an absence of some 10 years, spent traveling the Mediterranean, he returned to Athens, where he founded a school of philosophy near the grove dedicated to the early hero Acadēmos and hence known as the Academy. The select band of scholars who gathered there engaged in philosophical disputations in preparation for their role as leaders. Good government, Plato believed, would only come from an educated society in which kings are philosophers, and philosophers, kings.

Plato's literary dialogues provide a comprehensive picture of his approach to education. Basically, it was built around the study of dialectic (the skill of accurate verbal reasoning), which, if pursued properly, he believed, enables misconceptions and confusions to be stripped away and the nature of underlying truth to be established. The ultimate educational quest, as revealed in the dialogues, is the search for the Good, that is, the ultimate idea that binds together all earthly existence.

Plato's educational program is set out in his most famous dialogue, *The Republic.* The world, he argued, has two aspects, the visible, or that which is perceived with the senses, and the non-visible, or the intelligible, which consists of universal, eternal forms or ideas that are apprehensible only by the mind. Furthermore, the visible realm itself is subdivided into two, the realm of appearances and that of beliefs. Human experiences of so-called reality, according to Plato, are only of visible "appearances" and from these can be derived only opinions and beliefs. Most people, he argued, remain locked in this visible world of opinion; only a select few can cross into the realm of the intelligible. Through a rigorous 15-year program of higher education devoted to the study of dialectics and mathematical reasoning, this elite ("persons of gold" was Plato's term) can attain an understanding of genuine reality, which is composed of such forms as goodness, truth, beauty, and justice. Plato maintained that only those individuals who survive this program are really fit for the highest offices of the state and capable of being entrusted with the noblest of all tasks, those of maintaining and dispensing justice.

The education of Isocrates

The rival school of Isocrates was much more down-to-earth and practical. It too aimed at a form of wisdom but of a much more practical order, based on working out commonsense solutions to life's problems. In contrast to Plato, Isocrates sought to develop the quality of grace, cleverness, or finesse rather than the spirit of geometry. The program of study that he enjoined upon his pupils was more literary than scientific. In addition to gymnastics and music, its basics included the study of the Homeric classics and an extensive study of rhetoric—consisting of five or six years of theory, analysis of the great classics, imitation of the classics, and finally practical exercises.

These two parallel forms of culture and of higher education were not totally in conflict: both opposed the cynical pragmatism of the Sophists; each influenced the other. Isocrates did promote elementary mathematics as a kind of mental training or mental gymnastics and did allow for a smattering of philosophy to illumine broad questions of human life. Plato, for his part, recognized the usefulness of the literary art and philosophical rhetoric. The two traditions appear as two species of one genus; their debate, continued in each generation, enriched classical culture without jeopardizing its unity.

Aristotelian education

Before leaving the Hellenic age, there is one other great figure to appraise—one who was a bridge to the next age since he was the tutor of the young prince who became Alexander the Great of Macedonia. Aristotle (384–322 BC), who was one of Plato's pupils and shared some of his opinions about education, believed that education should be controlled by the state and that it should have as a main objective the training of citizens. The last book of his *Politics* opens with these words:

> No one will doubt that the legislator should direct his attention above all to the education of youth.... The citizen should be moulded to suit the form of government under which he lives.

He shared some of Plato's misgivings about democracy; but, because he was no recluse but a man of the world acquainted with public affairs, he declared his preference for limited democracy, "polity," over other forms of government. His worldliness also led him to be less concerned with the search for ideas, in the Platonic mode, and more concerned with the observation of specific things. His urge for logical structure and classification, for systematization, was especially strong.

This systematization extended to a youth's education. In his first phase, from birth to age seven, he was to be physically developed, learning how to endure hardship. From age seven to puberty, his curriculum would include the fundamentals of gymnastics, music, reading, writing, and enumeration. During the next phase, from puberty to age 17, the student would be more concerned with exact knowledge, not only carrying on with music and mathematics but also exploring grammar, literature, and geography. Finally, in young manhood, only a few superior students would continue into higher education, developing encyclopaedic and intensely intellectual interests in the biological and physical sciences, ethics, and rhetoric, as well as philosophy. Aristotle's school, the Lyceum, was thus much more empirical than Plato's Academy.

The Hellenistic age. Alexander the Great's conquest of the Persian empire between 334 and 323 BC abruptly extended the area of Greek civilization by carrying its eastern frontier from the shores of the Aegean to the banks of the Syrdarya and Indus rivers in Central Asia. Its unity rested henceforward not so much on nationality (it incorporated and assimilated Persians, Semites, and Egyptians) nor on the political unity soon broken after the death of Alexander in 323 but on a common Greek way of life, the fact of sharing the same conception of man. This ideal was no longer social, communal in character, as had been that of the city-state; it now concerned man as an individual—or, better, as a person. This civilization of the Hellenistic age has been defined as a civilization of paideia—which eventually denoted the condition of a person achieving enlightened, mature self-fulfillment but which originally

The concept of paideia

signified education per se. The Greeks succeeded in preserving their distinctive national way of life amid this immense empire because, wherever numbers of them settled, they brought with them their own system of education for their youth, and they not only resisted being absorbed by the "barbarian" non-Hellenic peoples but succeeded somewhat in spreading Greek culture to many of the alien elite. It is important to note that, although Hellenism was finally to be swept away in the Middle East by the Persian national renaissance and the invasions originating from Central Asia beginning in the 2nd century BC, it continued to flourish and even expand in the Mediterranean world under Roman domination. Hellenistic civilization and its educational pattern were prolonged to the end of antiquity and even beyond; it was to be a slow metamorphosis and not a brutal revolution that would later give birth to the civilization and education strictly called Byzantine.

The institutions. Hellenistic education comprised an ensemble of studies occupying the young from age seven to age 19 or 20. To be sure, this entire program was completed only by a minority, recruited from the rich aristocratic and urban bourgeois classes. The students were mostly boys (girls occupied only a very modest place), and of course they were usually free citizens (masters, though some slaves were given a professional education occasionally reaching a high level).

As in the preceding era, education continued to be dependent upon the city, which remained the primary frame of Greek life. To facilitate control of his empire, Alexander had commenced the process of founding a network of cities or communities organized and administered in the Greek manner. In effect, the creation of vast kingdoms did not eliminate the role of the city, even if the latter was not altogether independent; the Hellenistic state was not at all totalitarian and sought to reduce its administrative machinery to a minimum. It relied upon the cities to assume responsibility for public services, that of education in particular. The city in turn looked to the contributions of the richest and most generous private individuals, either by requiring them to fill magistracies and supply costly services or by appealing to their voluntary generosity; the proper functioning of the Hellenistic city presupposed the willing contributions of "benefactors." Thus, certain educational institutions were supported—and in fact sometimes set up—by private foundations that specified exactly the use to be made of the income from their gift of capital. Many schools were private, the role of the city being limited to inspections and to the organization of athletic and musical competitions and festivals.

Physical education. The Hellenistic school par excellence was still the school of gymnastics, the practice of athletic sports and the nudity that they required being the most characteristic feature contrasting the Greek way of life with that of the barbarians. There were, at least in sufficiently large cities, several gymnasiums, separately for the different age classes and on occasion for the sexes. They were essentially palaestrae, or open-air, square-shaped sports grounds, surrounded by colonnades in which were set up the necessary services: cloakrooms, washstands, training rooms, massage rooms, and classrooms. Outside there was a track for footraces, the *stadion.*

The gymnasium

The foundation of the training always consisted of the sports properly called gymnastic and field. Horsemanship remained an aristocratic privilege. Nautical sports had a very modest role—a curious thing for a nation of sailors, but the fact is the Greeks were by origin Indo-Europeans from the interior of the Eurasian continent. The other sports—ball games, hockey—were considered merely diversions or at best preparatory exercises. As the competition of professional sports grew, however, education based on sports progressively, though no doubt very slowly, lost its preeminent position. The popularity of athletic sports as spectacle endured, but educational sports moved into the background, disappearing altogether in the Christian period (in the 4th century AD) in favour of literary studies.

There was a similar progressive decline, a similar final effacement, of artistic, particularly musical, education, the other survivor from the Archaic culture. The art of music continued to flourish, but like sports it became the concern of professional practitioners and a feature of public spectacles rather than an art generally practiced in cultivated circles.

The primary school. The child from seven to 14 years of age went to the school of letters, conducted thither, as in the classical period, by the *paidagōgos,* whose role was not limited to accompanying the child: he had also to educate him in good manners and morals and finally to act as a lesson coach. Literacy and numeration were taught in the private school conducted by the *grammatistes.* Class sizes varied considerably, from a few pupils to perhaps dozens. The teaching of reading involved an analytical method that made the process very slow. First the alphabet was taught from alpha to omega, and then backward, then from both ends at once—alpha–omega, beta–psi, and so on to mu–nu. (A comparable progression in the Latin alphabet would be *A–Z, B–Y,* and so on to *M–N.*) Then were taught simple syllables—*ba, be, bi, bo*—followed by more complex ones, and then by words, successively of one, two, and three syllables. The vocabulary list included rare words (*e.g.,* some of medical origin), chosen for their difficulty of reading and pronunciation. It took several years for the child to be able to read connected texts, which were anthologies of famous passages. With reading was associated recitation and, of course, practice in writing, which followed the same gradual plan.

The program in mathematics was very limited; rather than computation, the subject, strictly speaking, was numeration: learning the whole numbers and fractions, their names, their written notations, their representation in finger counting (in assorted bent positions of the fingers and assorted placements of either hand relative to the body). The general use of tokens and of the abacus made the teaching of methods of computation less necessary than it became in the modern world.

Secondary education. Between the primary school and the various types of higher education, the Hellenistic educational system introduced a program of intermediate, preparatory studies—a preliminary education, a kind of common trunk preparing for the different branches of higher culture, *enkyklios paideia* ("general, or common, education"). This general education, far from having "encyclopaedic" ambitions in the modern sense of the word, represented a reaction against the inordinate ambitions of philosophy and, more generally, of the Aristotelian ideals of culture, which had demanded the large accumulation of intellectual attainments. The program of the *enkyklios paideia* was limited to the common points on which, as noted earlier, the rival pedagogies of Plato and of Isocrates agreed, namely, the study of literature and mathematics. Specialized teachers taught each of these subjects. The mathematics program had not changed since the ancient Pythagoreans and comprised four disciplines—arithmetic, geometry, astronomy, and harmonics (not the art of music but the theory of the numerical laws regulating intervals and rhythm). The primary function of the *grammatikos,* or professor of letters, was to present and explicate the great classic authors: Homer first of all, of whom every cultivated man was expected to have a deep knowledge, and Euripides and Menander—the other poets being scarcely known except through anthologies. Although poetry remained the basis of literary culture, room was made for prose—for the great historians, for the orators, Demosthenes in particular, even for the philosophers. Along with these explications of texts, the students were introduced to exercises in literary composition of a very elementary character (for example, summarizing a story in a few lines).

The program of this intermediate education did not attain its definitive formulation until the second half of the 1st century BC, after the appearance of the first manual devoted to the theoretical elements of language, a slim grammatical treatise by Dionysius Thrax. The program then consisted of the seven liberal arts: the three literary arts of grammar, rhetoric, and dialectic and the four mathematical disciplines noted above. (These were, respectively, the trivium and the quadrivium of medieval education, though the latter term did not appear until the 6th century and the former not until the 9th century.) The long career of this program should not conceal the fact that in

The seven liberal arts

the course of the centuries it fell into disuse and became rather largely a theory or abstraction; in reality, literary studies gradually took over at the expense of the sciences. Of the four mathematical disciplines, only one remained in favour—astronomy. And this was not merely because of its connections with astrology but primarily because of the popularity of the basic textbook used to teach it—the *Phaenomena,* a poem in 1,154 hexameters by Aratus of Soli—whose predominantly literary quality was suited to textual explications. Not until about the 3rd and 4th centuries AD was the need of a sound preparatory mathematical education again recognized and put into practice.

Higher education. Higher education appeared in several forms, complementary or competitive. First was the *ephebeia* ("youth" culture), a kind of civic and military training that completed the education of the young Greek and prepared him to enter into life; it lasted two years (from 18 to 20) and corresponded quite closely to the obligatory military service of modern states. It was a survival from the regime of the old Greek city-states, but in the Hellenistic age the absence of national independence erased all reason for this military training; between the 3rd and 2nd centuries BC the Athenian *ephebeia* (eventually reduced to a single year) was transformed into a leisured civilian college where a minority of rich young men came to be initiated into the refinements of the elegant life. Military training came to play only a modest role and gave way to athletic competition. To this were added lectures on scientific and literary subjects, assuring the ephebe a polish of general culture. The same evolution took place in other cities: the *ephebeia* became everywhere more aristocratic than civic, more sporting than military. What the Greeks, especially those who had emigrated to the barbarian lands, demanded of it was above all that it initiate their sons into Greek life and its characteristic customs, beginning with athletic sports. Especially in Egypt, it was intended to legitimize the privileged status of the Hellene relative to the "native" Egyptian. In any event, the *ephebeia* no longer was the setting for the highest forms of education.

Education in science and medicine

Formal education in science also lacked any institutionalization. There were, however, some establishments having scientific staffs of high competence, of which the most important was the Mouseion (Museum) established at Alexandria, richly endowed by the Ptolemies; but, at least initially, it was an institute for advanced research. If the scholars endowed there were also teachers, this meant only that they dispensed instruction to a small circle of chosen disciples. The same informal character of personal training was to be seen in all the special disciplines—medicine, for example, which saw such a fine development between the time of Hippocrates (5th century BC) and that of Galen (2nd century AD). If there were in the Hellenistic era certain "schools" of medicine—old (Cnidus, Cos) and new (Pergamum, Alexandria)—these were less the equivalent of today's medical faculties than simply centres to which the presence of numerous qualified masters attracted a large number of aspirants. Whatever theory these "students" were able to learn, they learned largely through self-training and practice, by associating themselves with a practicing physician whom they accompanied to the bedsides of patients, taking part in his consultations, profiting by his experience and advice.

Philosophy and rhetoric were subjects of education most highly institutionalized. Although philosophy was taught privately by individual masters-lecturers, who could be either itinerants or residents of one place, these teachers were well organized and, in groups, possessed a kind of institutional character. On the model of Plato's Academy, the new Athenian schools of philosophy—Aristotle's Lyceum, Epicurus' Garden, the Porch (stoa), which gave its name to the Stoics—were brotherhoods in which the posts in both teaching and administration were passed from generation to generation as a kind of heritage. It was in philosophy that the personalistic character of the Hellenistic era most clearly asserted itself, in contrast to the more communal idea of the preceding period; when philosophy turned to the problem of politics, for instance, it dealt less with the citizens of a republic and more with the sovereign king, his duties and character. The central problem was henceforth that of wisdom, of the purpose that man should set for himself in order to attain happiness, the supreme ideal. The teaching of philosophy was not entirely contemplative: it involved the disciple in an experience analogous to a religious conversion, a decision implying a revision of his life and the adoption of a generally ascetic way of life. Such a vocation, however, could obviously appeal only to a moral, intellectual, and financially secure elite; philosophers were always quite a small number within the Hellenistic (and Roman) intelligentsia.

The prestige of rhetoric

The reigning discipline was always rhetoric. The prestige of the oratorical art outlived those social conditions that had inspired it; political eloquence operated only in the context of an embassy coming to plead the cause of a particular city or pressure group at the court of the sovereign. Legal eloquence maintained its function, and the profession of advocate retained its attractiveness; but it was above all the eloquence of showy set speeches, the art of the lecturer, that experienced a curious blossoming. Also, as a result of the customary habit of reading aloud, there was no sharp line between speech and the book; thus, eloquence imposed its rule upon all literary genres—poetry, history, philosophy. Even the astronomer and the physician became lecturers.

Hence, great importance was attached to the teaching of rhetoric, which developed from century to century with an ever more rigorous technicalism, precision, and systematization. The study of rhetoric had five parts: invention (the art of finding ideas, according to standard schemes), disposition (the arrangement of words and sentences), elocution, mnemonics (memory training), and action. Action was the art of self-presentation, the regulation of voice and delivery, and above all the art of reinforcing the word with the expressive power of gesture. Each of these parts, equally systematized to the tiniest detail, was taught with a technical vocabulary of extreme precision. Such an education, which in addition to theory comprised a study of the great examples to be imitated and exercises in practical application, required many years of study; in fact, even in maturity, the cultivated Hellene continued to deepen his knowledge of the art, to drill himself, to "declaim."

A rivalry existed between philosophy and rhetoric, each trying to draw into its orbit the best and the most students. Even in the time of Plato and Isocrates, this rivalry did not proceed without mutual concessions and reciprocal influences, but it remained one of the most constant characteristics of the classical tradition and continued until the end of antiquity and beyond. The long summer of Hellenic civilization was extended under the Roman domination; the great centres of learning also experienced a long prosperity. Athens in particular was the unchallenged capital of philosophy; its *ephebeia* welcomed foreigners to come to crown their culture in the "school of Greece." Its masters of eloquence also had a solid reputation, even though they had competition from such schools of Asia Minor as those of Rhodes (in the 1st century BC) and Smyrna (in the 2nd century AD). Under the later Roman Empire, Alexandria, already famous for medicine, competed with Athens for preeminence in philosophy. Other great centres developed: Beirut, Antioch, and the new capital Constantinople. The quality of the teachers and the number of students attending permits one to apply to these centres, without too much anachronism, the modern designation of "universities," or institutions of advanced learning.

ANCIENT ROMANS

Early Roman education. The quality of Latin education before the 6th century BC can only be conjectured. Rome and Roman civilization were then dominated by a rural aristocracy of landed proprietors directly engaged in exploiting their lands, even after the establishment of the republic. Their spirit was far removed from Greece and Homeric chivalry; ancient Roman education was instead an education suitable for a rural, traditional people—instilling in youth an unquestioned respect for the customs of the ancestors: the *mos maiorum.*

Education had a practical aspect, involving instruction in such farm management concerns as how to oversee the work of slaves and how to advise tenant farmers or one's

steward. It had a legal aspect; in contrast to Athenian law, which relied more on common law than on codified law, Roman justice was much more formalistic and technical and demanded much more study on the part of the citizen. Education also had a moral aspect, aiming at inculcating rural virtues, a respect for good management of one's patrimony, and a sense of austerity and frugality. Roman education, however, did not remain narrowly utilitarian; it broadened in urban Rome, where there developed the same ideal of communal devotion to the public weal that had existed in Greece—with the difference that in Rome such devotion would never be called into question. The interests of the state constituted the supreme law. The ideal set before youth was not that of the chivalrous hero in the Homeric manner but that of the great men of history who, in difficult situations, had by their courage and their wisdom saved the fatherland when it was in danger. A nation of small farmers, Rome was also a nation of soldiers. Physical education was oriented not toward self-realization or competitive sport but toward military preparedness: training in arms, toughening of the body, swimming across cold and rapid streams, and horsemanship, involving such performances as mounted acrobatics and cavalry parades under arms.

The familial character of Roman education

Differing from the Greeks, the Romans considered the family the natural milieu in which the child should grow up and be educated. The role of the mother as educator extended beyond the early years and often had lifelong influence. If, in contrast to the girl, the boy at seven years of age was allowed to move away from her exclusive direction, he came under the control of his father; the Roman father closely supervised the development and the studies of his son, giving him instruction in an atmosphere of severity and moral exigency, through precept but even more through example. The young Roman noble accompanied his father as a kind of young page in all his appearances, even within the Senate.

Familial education ended at 16, when the adolescent male was allowed to wear adult dress, the pure white woolen toga virilis. He devoted one year to an apprenticeship in public life, no longer at his father's side but placed in the care of some old friend of the family, a man of politics laden with years and honours. Then came military service, first as a simple soldier (it was well for the future leader to learn first to obey), encountering his first opportunity to distinguish himself by courage in battle, but soon thereafter as a staff officer under some distinguished commander. Civil and military, the education of the young Roman was thus completed in the entourage of some high personage whom he regarded with respect and veneration, without ceasing, however, to gravitate toward the family orbit. The young Roman was brought up not only to respect the national tradition embodied in the example of the illustrious men of the past but also, very specifically, to respect the particular traditions of his own family, which too had had its great men and which jealously transmitted a stereotype, a specific attitude toward life. If ancient Greek education can be defined as the imitation of the Homeric hero, that of ancient Rome took the form of imitation of one's ancestors.

Roman adoption of Hellenistic education. Something of these original characteristics was to survive always in Roman society, so ready to be conservative; but Latin civilization did not long develop autonomously.

It assimilated, with a remarkable faculty for adaptation, the structures and techniques of the much further evolved Hellenistic civilization. The Romans themselves were quite aware of this, as evidenced by the famous lines of Horace: "Captive Greece captivated her rude conqueror and introduced the arts to rustic Latium" ("Graecia capta ferum victorem cepit et artis intulit agresti Latio" [*Epistles,* II, i, 156]).

Greek influence was felt very early in Roman education and grew ever stronger after the long series of gains leading to the annexation of Macedonia (168 BC), of Greece proper (146 BC), of the kingdom of Pergamum (133 BC), and finally of the whole of the Hellenized Orient. The Romans quickly appreciated the advantages they could draw from this more mature civilization, richer than their own national culture. The practical Romans grasped the advantages to be drawn from a knowledge of Greek, an international language known to many of their adversaries, soon to be their Oriental subjects, and grasped the related importance of mastering the art of oratory so highly developed by the Greeks. Second-century Rome assigned to the spoken word, particularly in political and legal life, as great an importance as had Athens in the 5th century. The Roman aristocrats quickly understood what a weapon rhetoric could be for a statesman.

Rome doubly adopted Hellenistic education: on the one hand, it came to pass that a Roman was considered truly cultivated only if he had the same education, in Greek, as a native Greek acquired; on the other hand, there progressively developed a parallel system of instruction that transposed into Latin the institutions, programs, and methods of Hellenistic education. Naturally, only the children of the ruling class had the privilege of receiving the complete and bilingual education. From the earliest years, the child, boy or girl, was entrusted to a Greek servant or slave and thus learned to speak Greek fluently even before being able to speak Latin competently; the child also learned to read and write in both languages, with Greek again coming first. (Alongside this private tutoring there soon developed, from the 3rd century BC, a Greek public education in schools aimed at a socially broader clientele, but the results of this schooling were less satisfactory than the direct method enjoyed by the children of the aristocracy.) In following the normal course of studies, the young Roman was taught next by an instructor of Greek letters (*grammatikos*) and then by a Greek rhetorician. Those desiring more complete training did not content themselves with the numerous and often highly qualified Greeks to be found in Rome itself but went to Greece to participate in the higher studies of the Greeks themselves. From 119 or 118 BC onward, the Romans secured admission to the Ephebic College at Athens, and in the 1st century BC such young Latins as Cicero were attending the schools of the best philosophers and rhetoricians at Athens and Rhodes.

The tutoring of Roman children

Roman modifications. The adoption of Hellenistic education did not proceed, however, without a certain adaptation to the Latin temperament: the Romans showed a marked reserve toward Greek athleticism, which shocked both their morals and their sense of the deep seriousness of life. Although gymnastic exercises entered into their daily life, it was under the category of health and not that of sport; in Roman architecture, the palaestra or gymnasium was only an appendage of the public baths, which were exaggerations of their Greek models. There was the same reserve, on grounds of moral seriousness, toward music and dance, arts suitable for professional performers but not for freeborn young men and least of all for young aristocrats. The musical arts indeed became integrated into Latin culture as elements of the life of luxury and refinement, but as spectacle rather than as amateur participation; hence their disappearance from programs of education. It must be remembered, however, that athletics and music were in Greece itself survivals of archaic education and had already entered upon a process of decline.

This education in a foreign language was paralleled by a course of studies exactly patterned upon those of the Greek schools but transposed into the Latin language. The aristocracy was to remain always attached to the idea of private education conducted within the family, but social pressure brought about the gradual development of public education in schools, as in Greece, at three levels—elementary, secondary, and higher; they appeared at different dates and in various historical contexts.

Roman primary and secondary education

Education of youth. The appearance of the first primary schools is difficult to date; but the use of writing from the 7th century BC implies the early existence of some kind of appropriate primary instruction. The Romans took their alphabet from the Etruscans, who had taken theirs from the Greeks, who had taken theirs from the Phoenicians. The early Romans quite naturally copied the pedagogy of the Hellenistic world: the same ignorance of psychology, the same strict and brutal discipline, the same analytical method characterized by slow progress—the alphabet (forward, backward, from both ends toward the middle),

the syllabary, isolated words, then short sentences (one-line moral maxims), finally continuous texts—the same method for writing, and the same numeration, rather than computation.

It was only between the 3rd and the end of the 1st century BC that Latin secondary education developed, staffed by the *grammaticus Latinus,* corresponding to the Greek *grammatikos.* Since the principal object of this education was the explication of poetry, its rise was hindered by the slowness with which Latin literature developed. The first-known of these teachers, Livius Andronicus, took as his subject matter his own Latin translation of the *Odyssey;* two generations later, Ennius explicated his own poetic works. Only with the great poets of the age of Augustus could Latin literature provide classics able to rival Homer in educational value; they were adopted as basic texts almost immediately after their appearance. Thereafter, and until the end of antiquity, the program was not to undergo further change, the principal authors being first of all Virgil, the comic author Terence, the historian Sallust, and the unchallenged master of prose, Cicero. The methods of the Latin grammarian were copied directly from those of his Greek counterpart; the essential point was the explication of the classic authors, completed by a theoretical study of good language using a grammar textbook and by practical exercises in composition, graduated according to a minutely regulated progression and always remaining rather elementary. Theoretically, the curriculum remained that of the seven liberal arts, but, as in Greece, it practically neglected the study of the sciences in favour of that of letters.

Latin rhetoric

It was only in the 1st century BC that the teaching of rhetoric in Latin was established: the first recorded Latin rhetorician, Plotius Gallus, appeared in 93 BC in a political context, namely, as a democratic initiative to counter the aristocratic education given in Greek, and, as such, was soon prohibited by the conservative party in power. It was not until the end of the century and the appearance of the works of Cicero that this education would be revived and become normal practice; first, Cicero's discourses offered the young Latin the equivalent of those of the Greek Demosthenes, and, second, Cicero's theoretical treatises provided a technical vocabulary obviating the need for Greek manuals. But this instruction was to remain always very close to its Hellenistic origins: the terminology used by Rome's greatest educator, Quintilian (*c.* AD 35–*c.* 100), is much more impregnated with Hellenism, much less Latinized, than that which Cicero had proposed. At Rome, too, rhetoric became the form of higher education enjoying the greatest prestige; as in Greece, this popularity outlived the elimination of political eloquence. More than in Greece, legal eloquence continued to flourish (Quintilian had in mind particularly the training of future advocates), but, as in the Hellenic milieu, Latin culture became predominantly aesthetic: from the beginning of the empire, the public lecture was the most fashionable literary genre, and the teaching of rhetoric was very naturally oriented toward the art of the lecturer as the crowning achievement.

Higher education. Because the oratorical art was incontestably the most popular subject of higher education, the Romans did not feel the same urgency to Latinize the other rival branches of knowledge, which interested only a small number of specialists with unusual vocations. To be sure, the philosophical work of Cicero had the same ambition as his oratorical work and proved by its existence that it was possible to philosophize in Latin, but philosophy found no successors to Cicero as rhetoric did. There was never a Latin school for philosophy. Of course, Rome did not lack philosophers, but many used Greek as their means of expression (even the emperor Marcus Aurelius); those who, like Cicero, wrote in Latin—Seneca, for example—had taken their philosophy studies in Greek. It was the same in the sciences, particularly in the medical sciences; for long, there were no medical books in Latin except encyclopaedias on a popular level.

The innovation of legal education

On the other hand, Rome created in the school of law another type of higher education—the only one that had no equivalent in Hellenistic education. The position of law in Roman life and civilization is, of course, well known. Perhaps even more than rhetoric, it offered young Romans profitable careers; very naturally, there developed an appropriate education to prepare them. At first elementary in character and entirely practical, it was given within the framework of apprenticeship: the professor of law (*magister juris*) was primarily a practitioner, who initiated into his art the group of young disciples entrusted to him; these listened to his consultations and heard him plead or judge. Beginning in Cicero's time and undoubtedly under his influence, this instruction was paralleled by a systematic theoretical exposition. Roman law was thus promoted to the rank of a scientific discipline. True schools were progressively established and took on an official character; their existence is well attested beginning with the 2nd century AD. It was at this same time that legal education acquired its definitive tools, with the composition of systematic elementary treatises such as the *Institutiones* of Gaius, manuals of procedure, commentaries on the law, and systematic collections of jurisprudence. This creative period perhaps reached its peak at the beginning of the 3rd century AD. The works of the great legal authors of this time, which became classics, were offered by the law professor with much interpretation and explication—very similar to the way in which grammarians offered literature.

Rome, the capital, remained the great centre of this advanced study in law. At the beginning of the 3rd century, however, there appeared in the Roman Orient the school of Beirut. The teaching there was in Latin; and, to hear it and profit by the advantages that it offered for a high administrative or judicial career, many young Greeks enrolled at the school, in spite of the language obstacle. Only a legal career could persuade the Greeks to learn Latin, a language that they had always regarded as "barbarous."

The Roman world became covered with a network of schools concurrent with the Romanization of the provinces. The primary school always remained private; on the other hand, many schools of grammar or rhetoric acquired the character of public institutions supported (as in the Hellenic world) either by private foundations or by a municipal budget. In effect, it was always the city that was responsible for education. The liberal central government of the high empire, anxious to reduce its administrative apparatus to a minimum, made no pretense of assuming charge of it. It was content to encourage education and to favour teaching careers by fiscal exemptions; and only very exceptionally did an emperor create certain chairs of higher education and assign them a regular stipend. Vespasian (AD 69–79) created two chairs at Rome, one of Greek rhetoric and the other of Latin rhetoric. Marcus Aurelius (AD 161–180) similarly endowed, in Athens, a chair of rhetoric and four chairs of philosophy, one for each of the four great sects—Platonism, Aristotelianism, Epicureanism, and Stoicism.

The durable character of Greco-Roman education

Education in the later Roman Empire. The dominant fact is the extraordinary continuity of the methods of Roman education throughout such a long succession of centuries. Whatever the profound transformations in the Roman world politically, economically, and socially, the same educational institutions, the same pedagogical methods, the same curricula were perpetuated without great change for 1,000 years in Greek and six or seven centuries in Roman territory. At most, a few nuances of change need be noted. There was a measure of increasing intervention by the central government, but this was primarily to remind the municipalities of their educational duties, to fix the remuneration of teachers, and to supervise their selection. Only higher education received direct attention: in AD 425, Theodosius II created an institute of higher education in the new capital of Constantinople and endowed it with 31 chairs for the teaching of letters, rhetoric (both Greek and Latin), philosophy, and law. Another innovation was that the exuberant growth of the bureaucratic apparatus under the later empire favoured the rise of one branch of technical education, that of stenography.

The only evolution of any notable extent involves the use of Greek and Latin. There had never been more than a few Greeks who learned Latin, even though the growing machinery of administration and the increasing clientele drawn to the law schools of Beirut and Constantinople

tended to increase the numerical size of this tiny minority. On the other hand, in Latin territory, late antiquity exhibited a general recession in the use of Greek. Although the ideal remained unchanged and high culture always proposed to be bilingual, most people generally knew Greek less and less well. This retrogression need not be interpreted solely as a phenomenon of decadence: it had also a positive aspect, being an effect of the development of Latin culture itself. The richness and worth of the Latin classics explain why the youth of the West had less time than formerly to devote to the study of the Greek authors. Virgil and Cicero had replaced Homer and Demosthenes, just as in modern Europe the ancient languages have retreated before the progress of the national languages and literatures. Hence, in the later empire there appeared specialists in intercultural relations and translations from Greek into Latin. In the 4th and particularly in the 5th century, medical education in Latin became possible, thanks to the appearance of a whole medical (and veterinary) literature consisting essentially of translations of Greek manuals. It was the same with philosophy: resuming Cicero's enterprise at a distance of more than five centuries, Boethius (*c.* 480–524) in his turn sought with his manuals and his translations to make the study of that discipline available in Latin. Although the misfortunes of Italy in the 6th century, including the Lombardian invasion, did not permit this hope to be realized, the work of Boethius later nourished the medieval renaissance of philosophic thought.

Christian use of Greco-Roman education

Nothing better demonstrates the prestige and the allure of classical culture than the attitude taken toward it by the Christians. This new religion could have organized an original system of education analogous to that of the rabbinical school—that is, one in which children learned through study of the Holy Scriptures—but it did not do so. Usually, Christians were content to have both their special religious education, provided by the church and the family, and their classical instruction, received in the schools and shared with the pagans. Thus, they maintained the tradition of the empire after it had become Christian. Certainly, in their view, the education dispensed by these schools must have presented many dangers, inasmuch as classical culture was bound up with its pagan past (at the beginning of the 3rd century the profession of schoolteacher was among those that disqualified one from baptism); but the utility of classical culture was so evident that they considered it necessary to send their children to these same schools in which they barred themselves from teaching. From Tertullian to St. Basil the Great of Caesarea, Christian scholars were ever mindful of the dangers presented by the study of the classics, the idolatry and immorality that they promoted; nevertheless, they sought to show how the Christian could make good use of them.

With the passage of time and the general conversion of Roman society and particularly of its ruling class, Christianity, overcoming its reserve, completely assimilated and took over classical education. In the 4th century Christians were occupying teaching positions at all levels, from schoolmasters and grammarians to the highest chairs of eloquence. In his treatise *De doctrina Christiana* (426), St. Augustine formulated the theory of this new Christian culture: being a religion of the Book, Christianity required a certain level of literacy and literary understanding; the explication of the Bible required the methods of the grammarian; preaching a new field of action required rhetoric; theology required the equipment of philosophy. The synthesis of Christianity and classical education had become so intimate that, when the "barbarian" invasions swept away the traditional school along with many other imperial and Roman institutions, the church, needing a literary culture for the education of its clergy, kept alive the cultural tradition that Rome had received from the Hellenistic world. (H.-I.M./J.Bo.)

Education in Persian, Byzantine, early Russian, and Islāmic civilizations

ANCIENT PERSIA

The ancient Persian empire began when Cyrus II the Great initiated his conquests in 559 BC, and it ended when it was overrun by the Muslims in AD 651. Three elements dominated this ancient Persian civilization: (1) a rigorous and challenging physical environment, (2) the activist and positive Zoroastrian religion and ethics, and (3) a militant, expansionist people. These elements developed in the Persians an adventurous personality mingled with intense national feelings.

Zoroastrian influences

In the early period (559–330 BC), known as the Achaemenid period for the dynastic name of Cyrus and his successors, education, sustained by Zoroastrian ethics and the requirements of a military society, aimed at serving the needs of four social classes—priests, warriors, tillers of the soil, and merchants. Three principles sustained Zoroastrian ethics: the development of good thoughts, of good words, and of good actions (see ZOROASTRIANISM AND PARSIISM). Achaemenid-Zoroastrian education stressed strong family ties and community feelings, acceptance of imperial authority, religious indoctrination, and military discipline.

Education was a private enterprise. Formative education was carried on in the home and continued after the age of seven in court schools for children of the upper classes. Secondary and higher education included training in law to prepare for government service, as well as medicine, arithmetic, geography, music, and astronomy. There were also special military schools.

In 330 BC Persia was conquered by Alexander the Great, and native Persian or Zoroastrian education was largely eclipsed by Hellenistic. Greek practices continued during the Parthian empire (247 BC–AD 224), founded by seminomadic conquerors from the Caspian steppes. And, thus, truly Persian influences were not restored until the appearance of a new, more sophisticated and reform-minded dynasty, the Sāsānians, in the 3rd century AD. In what has been called the neo-Persian empire of the Sāsānians (AD 224–651), the Achaemenid social structure and education were revived and further developed and modified. Zoroastrian ethics, though more advanced than during the Achaemenid period, emphasized similar moral principles but with new stress upon the necessity for labour (particularly agriculture), upon the sanctity of marriage and family devotion, and upon the cultivation of respect for law and of intellectualism—all giving to education a strong moral, social, and national foundation. The subject matter of basic education included physical and military exercises, reading (Pahlavi alphabet), writing (on wooden tablets), arithmetic, and the fine arts.

The Academy of Gondēshāpūr

The greatest achievement of Sāsānian education was in higher education, particularly as it developed in the Academy of Gondēshāpūr. Here, Zoroastrian culture, Indian and Greek sciences, Alexandrian-Syrian thought, medical training, theology, philosophy, and other disciplines developed to a high degree, making Gondēshāpūr the most advanced academic centre of learning in the later period of Sāsānian civilization. The academy, to which came students from various parts of the world, advanced, among other subjects, Zoroastrian, Greek, and Indian philosophies; Persian, Hellenic, and Indian astronomy; Zoroastrian ethics, theology, and religion; law, government, and finance; and various branches of medicine.

It was partly through the Academy of Gondēshāpūr that important elements of classical Greek and Roman learning reached the Muslims during the 8th and 9th centuries AD and through them, in Latin translations of Arabic works, the Schoolmen of western Europe during the 12th and 13th centuries. (M.K.N.)

THE BYZANTINE EMPIRE

The Byzantine Empire was a continuation of the Roman Empire in the eastern Mediterranean area after the loss of the western provinces to Germanic kingdoms in the 5th century. Although it lost some of its eastern lands to the Muslims in the 7th century, the empire lasted until Constantinople—the new capital founded by the Roman emperor Constantine the Great in 330—fell to the Ottoman Turks in 1453. The empire was seriously weakened in 1204 when, as a result of the Fourth Crusade, its lands were partitioned and Constantinople captured; but until then it remained a powerful centralized state, with a

common Christian faith, an efficient administration, and a shared Greek culture. This culture, already Christianized in the 4th and 5th centuries, was maintained and transmitted by an educational system that was inherited from the Greco-Roman past and based on the study and imitation of classical Greek literature.

Stages of education. There were three stages of education. The basic skills of reading and writing were taught by the elementary-school master, or *grammatistes,* whose pupils generally ranged from six or seven to 10 years of age. The secondary-school master, or *grammatikos,* supervised the study and appreciation of classical literature and of literary Greek, from which the spoken Greek of everyday life differed more and more in the course of time, and Latin (until the 6th century). His pupils ranged in age from 10 to 15 or 16. Next, the rhetorician, or *rhētor,* taught pupils how to express themselves with clarity, elegance, and persuasiveness, in imitation of classical models. Speaking style was deemed more important than content or original thinking. An optional fourth stage was provided by the teacher of philosophy, who introduced pupils to some of the topics of ancient philosophy, often by reading and discussing works of Plato or Aristotle. Rhetoric and philosophy formed the main content of higher education.

Availability of education

Elementary education was widely available throughout most of the empire's existence, not only in towns but occasionally in the countryside as well. Literacy was therefore much more widespread than in western Europe, at least until the 12th century. Secondary education was confined to the larger cities. Pupils desiring higher education had almost always to go to Constantinople, which became the cultural centre of the empire after the loss to the Muslim Arabs of Syria, Palestine, and Egypt in the 7th century. Monasteries sometimes had schools in which young novices were educated, but they did not teach lay pupils. Girls did not normally attend schools, but the daughters of the upper classes were often educated by private tutors. Many women were literate, and some, such as the hymnographer Kasia (9th century) and the historian-princess Anna Comnena (1083–*c.* 1153), were recognized as writers of distinction.

Elementary education. Elementary-school pupils were taught to read and write individual letters first, then syllables, and finally short texts, often passages from the Psalms. They probably also learned simple arithmetic at this stage. Teachers had a humble social status and depended on the fees paid by parents for their livelihood. They usually held classes in their own homes or on church porches but were sometimes employed as private tutors by wealthy households. They had no assistants and used no textbooks. Teaching methods emphasized memorization and copying exercises, reinforced by rewards and punishments.

Secondary education. The secondary-school teacher taught the grammar and vocabulary of classical and ecclesiastical Greek literature from the Hellenistic and Roman periods and explained the elements of classical mythology and history that were necessary for the study of a limited selection of ancient Greek texts, mainly poetry, beginning with Homer. The most commonly used textbook was the brief grammar by Dionysius Thrax; numerous and repetitive later commentaries on the book were also frequently used. From the 9th century on, these books were sometimes supplemented with the *Canons* of Theognostos, a collection of brief rules of orthography and grammar. The *grammatikos* might also make use of anonymous texts dating from late antiquity, which offered word-by-word grammatical explanations of Homer's *Iliad,* or of similar texts on the Psalms by Georgius Choiroboscos (early 9th century). Pupils would not normally possess copies of these textbooks, since handwritten books were very expensive, but would learn the rules by rote from their teacher's dictation. Beginning in the 11th century, much use was made in secondary education of *schedē* (literally, "sketches" or "improvisations"), short prose texts that often ended in a few lines of verse. These were specially written by a teacher to illustrate points of grammar or style. From the early 14th century on, much use was also made of *erotemata,* systematic collections of questions and answers on grammar which the pupil learned by heart.

Secondary schools often had more than one teacher, and the older pupils were often expected to help teach their juniors. Schools of this kind had little institutional continuity, however. The most lasting schools were those conducted in churches.

Higher education. The rhetorician's textbooks included systematic handbooks of the art of rhetoric, model texts with detailed commentaries, and specimens of oratory by classical or postclassical Greek writers and by Church Fathers, in particular Gregory of Nazianzus. Many Byzantine handbooks of rhetoric survive from all periods. They are often anonymous and always derivative, mostly based directly or indirectly on the treatises of Hermogenes of Tarsus (late 2nd century AD). There is little innovation in the theory of rhetoric that they expound. After studying models, pupils went on to compose and deliver speeches on various general topics.

Until the early 6th century there was a flourishing school of Neoplatonic philosophy in Athens, but it was repressed or abolished in 529 because of the active paganism of its professors. A similar, but Christian, school in Alexandria survived until the Arab conquest of Egypt in 640. For the next five centuries philosophical teaching seems to have been limited to simple surveys of Aristotle's logic, but in the 11th century there was a renewal of interest in the Greek philosophical tradition and many commentaries on works of Aristotle were composed, evidently for use in teaching. In the early 15th century the philosopher George Gemistos Plethon revived interest in Plato, who until then had been neglected for Aristotle. All philosophical teaching in the Byzantine world was concerned with the explanation of texts rather than with the analysis of problems.

State and church patronage of education

Because higher education provided learned and articulate personnel for the sophisticated bureaucracies of state and church, it was often supported and controlled officially, although private education always existed as well. There were officially appointed teachers in Constantinople in the 4th century, and in 425 the emperor Theodosius II established professorships of Greek and Latin grammar, rhetoric, and philosophy, but these probably did not survive the great crisis of the Arab and Slav invasions of the 7th century. In the 9th century the School of Magnaura, an institution of higher learning, was founded by imperial decree. In the 11th century Constantine IX established new schools of philosophy and law at the Capitol School in Constantinople. Both survived until the 12th century, when the school under the control of the patriarch of Constantinople, with teachers of grammar, rhetoric, and biblical studies, gained predominance. After the interval of Western rule in Constantinople (1204–61), both emperors and patriarchs gave sporadic support to higher education in the capital. As the power, wealth, and territory of the empire were eroded in the 14th and 15th centuries, the church became the principal and ultimately the only patron of higher education.

Professional education. Teaching of such professional subjects as medicine, law, and architecture was largely a matter of apprenticeship, although at various times there was some imperially supported or institutionalized teaching.

Strangely, there is little sign of systematic teaching of theology, apart from that given by the professors of biblical studies in the 12th-century patriarchal school. Studious reading of works by the Church Fathers was the principal path to theological knowledge in Byzantium, both for clergy and for laymen. Nonetheless, religious orthodoxy, or faith, was Byzantium's greatest strength. It held the empire together for more than 1,000 years against eastern invaders. Faith was also the Byzantine culture's chief limitation, choking originality in the sciences and the practical arts. But within this limitation it preserved the literature, science, and philosophy of classical Greece in recopied texts, some of which escaped the plunders of the crusaders and were carried to southern Italy, restoring Greek learning there. Combined with the treasures of classical learning that reached Europe through the Muslims, this Byzantine heritage helped to initiate the beginnings of the European Renaissance. (M.K.N./R.B.)

EARLY RUSSIAN EDUCATION: KIEV AND MUSCOVY

Properly, the term Russia applies only to the empire that covered roughly the present area of the Soviet Union from the 18th to the early 20th century. It is sometimes less strictly employed, however, as in this section, to refer to that area from ancient times as well.

Early influences on Russian education and culture

The influences of the Byzantine Empire and of the Eastern Orthodox church made themselves strongly felt in Russia as early as the 10th century, when Kiev, the first East Slavic state, was firmly established. At that time Prince Svyatoslav, a determined pagan, failed to maintain control of the route "from the Varangians to the Greeks" (south from Novgorod through Kiev, along the Dnepr River) and the Byzantine Empire expelled him from its Balkan possessions, which he was attempting to conquer. After his death in 972 the way lay open for sustained penetration of cultural influences emanating from Byzantium into the Kievan state, although formal relations between the two powers were seldom harmonious. Byzantine cultural materials entering the Kievan state were translated into Old Church Slavonic; thus, there was no language barrier. A famous tale in an early chronicle recounts how Grand Prince Vladimir in 988 ordered the people of Kiev to receive baptism in the Orthodox Christian rite. It is, however, highly dubious to claim that this event, which established Christianity as the predominant cultural force in the Kievan state, also marked the beginning of an institutionalized system of education. A few sources of the time spoke of "book learning," but all this actually meant was that people were expected to be acquainted with the rudiments of Holy Writ.

The next epoch in Russian history is known as the appanage period. This period runs roughly from the decline of Kiev in the 11th century to the rise of the grand principality of Moscow (Muscovy) in the 14th century. It was characterized by the appearance of numerous autonomous fiefdoms and a population shift from southern plains to northern forests, brought about in large part by attacks from steppe nomads. Although the church and monasteries continued to acquire wealth and property, anarchic decentralization was not conducive to the development of any kind of widespread, uniform educational apparatus.

During this time of instability, in 1240, the Mongol (or Tatar) empire, known as the Golden Horde, sacked and devastated the European Russian Plain and imposed its control over the region, although with diminishing efficiency, until 1451. The Mongol rule had a debilitating effect on all phases of Russian culture, including the church, which became more formalistic and ritualistic. What little can be learned about education at this time must be culled from later biographies of contemporary saints. It is not clear who served as teachers, how many there were, where they taught, or how many and what kind of pupils they had. What instruction they gave was of an uncompromisingly religious nature: seven-year-olds did little more than read aloud and chant devotional materials or, very rarely, recite the numbers from one to 100. Because students uttered their assignments simultaneously, the result was often chaotic.

By the time the Mongol rule came to an end, the welter of independent Russian principalities had been united under the authority of the grand principality of Moscow, which began a successful program of territorial expansion. Controversies over religious issues, particularly the respective roles of church and state, flared up but failed to bring about any real improvement in education. The church's inability to provide adequate education was recognized, however, and in 1551 a church council known as the Hundred Chapters was convened at the initiative of the tsar Ivan IV the Terrible. The council heard many stories of clerical ignorance and licentiousness, and its deliberations made it clear that no effective system or institution existed to educate the clergy, the key class in the cultural establishment.

The familial character of Russian education

It is misleading to think of education solely in institutional terms, however. Another system existed in early Russia: the highly developed family system, within which from generation to generation parents handed on to their children skills and knowledge. Indeed, the very strength and tenacity of the family unit may well have retarded development of a more formal educational structure.

Things began to change in the 17th century. It is necessary to bear in mind that Kiev and much of the western Ukraine had for centuries been under the control of the Roman Catholic Polish-Lithuanian state, where intellectual achievement and ferment, especially during the Renaissance and Reformation, had been considerably greater than in Muscovite Russia. The people of the Ukraine were determined to preserve Orthodoxy from Catholic pressure, which grew intense when the Jesuits employed their excellent schools as one means to spearhead the Counter-Reformation. Different Orthodox groups responded to the challenge by forming schools at many levels, culminating in the foundation of the Kievan Academy by Peter Mogila, the energetic metropolitan of Kiev, who strove to adapt Western educational techniques to defend Orthodoxy. It should be noted, however, that, although these schools adopted portions of the broader Western curriculum, their goal continued to be what it always had been, the inculcation of traditional religious values.

By the mid-17th century much of the western Ukraine had come under Muscovite control, enabling a number of educated Ukrainians, some trained in Poland, a few even in Rome, to come to Moscow. They arrived under the auspices of Patriarch Nikon, who was then engaged in correcting what he saw as errors in Orthodox church books; but their appearance aroused deep suspicion on the part of the Orthodox establishment, many of whose members displayed little interest in or sympathy for the establishment of schools, an undertaking the newcomers considered to be of primary importance. Educational reforms nevertheless continued, albeit slowly.

The reign of Peter I the Great (1682–1725) ushered in a new and more dynamic age, although even this ruler's reforming zeal proved inadequate to the central task of creating a national school system, particularly at the elementary level. Religion was deemphasized as Peter strove to establish at least a few institutions that would provide graduates trained in practical subjects for government and military service. Church schools were brought under state control, and the Academy of Sciences was established. Nevertheless, the creation of a network of schools capable at all levels of responding to Russia's rapidly changing priorities was a task that awaited the future. (H.F.Gr.)

THE ISLĀMIC ERA

Influences on Muslim education and culture. The Greco-Byzantine heritage of learning that was preserved through the medium of Middle Eastern scholarship was combined with elements of Persian and Indian thought and taken over and enriched by the Muslims. It was initiated as early as the Umayyad caliphate (661–750), which allowed the sciences of the Hellenistic world to flourish in Syria and patronized Semitic and Persian schools in Alexandria, Beirut, Gondēshāpūr, Nisibis, Haran, and Antioch. But the largest share of Islām's preservation of classical culture was assumed by the ʿAbbāsid caliphate (750–*c.* 1100), which followed the Umayyad and encouraged and supported the translation of Greek works into Arabic, often by Nestorian, Hebrew, and Persian scholars. These translations included works by Plato and Aristotle, Hippocrates, Galen, Dioscorides, Alexander of Aphrodisias, Ptolemy, and others. The great mathematician al-Khwārizmī (flourished 9th century) compiled astronomical tables, introduced Hindu numerals (which became Arabic numerals), formulated the oldest known trigonometric tables, and prepared a geographic encyclopaedia in cooperation with 69 other scholars.

The Muslim blend of scholastic heritages

The transmission of classical culture through Muslim channels can be divided into seven basic types: works translated directly from Greek into Arabic; works translated into Pahlavi, including Indian, Greek, Syriac, Hellenistic, Hebrew, and Zoroastrian materials, with the Academy of Gondēshāpūr as the centre of such scholarship (the works then being translated from Pahlavi into Arabic); works translated from Hindi into Pahlavi, then into Syriac, Hebrew, and Arabic; works written by Muslim scholars from the 9th through the 11th centuries but borrowed, in effect,

from non-Muslim sources, with the line of transmission obscure; works that amounted to summaries and commentaries of Greco-Persian materials; works by Muslim scholars that were advances over pre-Islāmic learning but that might not have developed in Islām had there not been the stimulation from Hellenistic, Byzantine, Zoroastrian, and Hindu learning; and, finally, works that appear to have arisen from purely individual genius and national cultures and would likely have developed independent of Islām's classical heritage of learning.

Aims and purposes of Muslim education. Islām placed a high value on education, and, as the faith spread among diverse peoples, education became an important channel through which to create a universal and cohesive social order. By the middle of the 9th century knowledge was divided into three categories: the Islāmic sciences, the philosophical and natural sciences (Greek knowledge), and the literary arts. The Islāmic sciences, which emphasized the study of the Qur'ān (the Islāmic scripture) and the Ḥadīth (the sayings and traditions of the Prophet Muḥammad) and their interpretation by leading scholars and theologians, were valued the most highly, but Greek scholarship was considered equally important albeit less virtuous.

Early Muslim education emphasized practical studies, such as the application of technological expertise to the development of irrigation systems, architectural innovations, textiles, iron and steel products, earthenware, and leather products; the manufacture of paper and gunpowder; the advancement of commerce; and the maintenance of a merchant marine. After the 11th century, however, denominational interests dominated higher learning, and the Islāmic sciences achieved preeminence. Greek knowledge was studied in private, if at all, and the literary arts diminished in significance as educational policies encouraging academic freedom and new learning were replaced by a closed system characterized by an intolerance toward scientific innovations, secular subjects, and creative scholarship. This denominational system spread throughout eastern Islām from Transoxania (roughly modern Uzbek S.S.R.) to Egypt, with some 75 schools in existence between about 1050 and 1250.

Organization of education. The system of education in the Muslim world was unintegrated and undifferentiated. Learning took place in a variety of institutions, among them the *ḥalqah,* or study circle; the *maktab* (*kuttab*), or elementary school; the palace schools; bookshops and literary salons; and the various types of colleges, the *meshed,* the *masjid,* and the madrasah. All the schools taught essentially the same subjects.

Varieties of Islāmic schools

The simplest type of early Muslim education was offered in the mosques, where scholars who had congregated to discuss the Qur'ān began, before long, to teach the religious sciences to interested adults. Mosques increased in number under the caliphs, particularly the 'Abbāsids: 3,000 of them were reported in Baghdad alone in the first decades of the 10th century; as many as 12,000 were reported in Alexandria in the 14th century, most of them with schools attached. Some mosques, such as that of al-Manṣūr, built during the reign of Hārūn ar-Rashīd in Baghdad, or those in Isfahan, Mashhad, Ghom, Damascus, Cairo, and the Alhambra (Granada), became centres of learning for students from all over the Muslim world. Each mosque usually contained several study circles (*ḥalqah*), so named because the teacher was, as a rule, seated on a dais or cushion with the pupils gathered in a semicircle before him. The more advanced a student, the closer he was seated to the teacher. The mosque circles varied in approach, course content, size, and quality of teaching, but the method of instruction usually emphasized lectures and memorization. Teachers were as a rule looked upon as masters of scholarship, and their lectures were meticulously recorded in notebooks. Students often made long journeys to join the circle of a great teacher. Some circles, especially those in which the Ḥadīth was studied, were so large that it was necessary for assistants to repeat the lecture so that every student could hear and record it.

Elementary schools (*maktab,* or *kuttab*), in which pupils learned to read and write, date to the pre-Islāmic period in the Arab world. After the advent of Islām, these schools developed into centres for instruction in elementary Islāmic subjects. Students were expected to memorize the Qur'ān as perfectly as possible. Some schools also included in their curriculum the study of poetry, elementary arithmetic, penmanship, ethics (manners), and elementary grammar. *Maktab*s were quite common in almost every town or village in the Middle East, Africa, Sicily, and Spain.

Schools conducted in royal palaces taught not only the curriculum of the *maktab*s but also social and cultural studies designed to prepare the pupil for higher education, for service in the government of the caliphs, or for polite society. The instructors were called *mu'addib*s, or instructors in good manners. The exact content of the curriculum was specified by the ruler, but oratory, history, tradition, formal ethics, poetry, and the art of good conversation were often included. Instruction usually continued long after the pupils had passed elementary age.

The high degree of learning and scholarship in Islām, particularly during the 'Abbāsid period in eastern Islām and the later Umayyads in western Islām, encouraged the development of bookshops, copyists, and book dealers in large, important Islāmic cities such as Damascus, Baghdad, and Córdoba. Scholars and students spent many hours in these bookshop schools browsing, examining, and studying available books or purchasing favourite selections for their private libraries. Book dealers traveled to famous bookstores in search of rare manuscripts for purchase and resale to collectors and scholars and thus contributed to the spread of learning. Many such manuscripts found their way to private libraries of famous Muslim scholars such as Avicenna, al-Ghazālī, and al-Fārābī, who in turn made their homes centres of scholarly pursuits for their favourite students.

Fundamental to Muslim education as were the circle schools, the *maktab*s, and the palace schools, they embodied definite educational limitations. Their curricula were limited; they could not always attract well-trained teachers; physical facilities were not always conducive to a congenial educational environment; and conflicts between religious and secular aims in these schools were almost irreconcilable. Most importantly, these schools could not meet the growing need for trained personnel or provide sufficient educational opportunities for those who wished to continue their studies. These pressures led to the creation of a new type of school, the madrasah, which became the crown and glory of medieval Muslim education. The madrasah was an outgrowth of the *masjid,* a type of mosque college dating to the 8th century. The differences between these two institutions are still being studied, but most scholars believe that the *masjid* was also a place of worship and that, unlike the madrasah, its endowment supported only the faculty and not the students as well. A third type of college, the *meshed* (shrine college), was usually a madrasah built next to a pilgrimage centre. Whatever their particularities, all three types of college specialized in legal instruction, each turning out experts in one of the four schools of Sunnite, or orthodox, Islāmic law.

Madrasahs may have existed as early as the 9th century, but the most famous one was founded in 1057 by the vizier Niẓām al-Mulk in Baghdad. The Niẓāmīyah, devoted to Sunnite learning, served as a model for the establishment of an extensive network of such institutions throughout the eastern Islāmic world, especially in Cairo, which had 75 madrasahs, in Damascus, which had 51, and in Aleppo, where the number of madrasahs rose from six to 44 between 1155 and 1260.

The great Muslim institutions

Important institutions also developed in the Spanish cities of Córdoba, Seville, Toledo, Granada, Murcia, Almería, Valencia, and Cádiz, in western Islām, under the Umayyads. The madrasahs had no standard curriculum; the founder of each school determined the specific courses that would be taught, but they generally offered instruction in both the religious sciences and the physical sciences.

The contribution of these institutions to the advancement of knowledge was vast. Muslim scholars calculated the angle of the ecliptic; measured the size of the Earth; calculated the precession of the equinoxes; explained, in the field of optics and physics, such phenomena as refraction of light, gravity, capillary attraction, and twilight; and

developed observatories for the empirical study of heavenly bodies. They made advances in the uses of drugs, herbs, and foods for medication; established hospitals with a system of interns and externs; discovered causes of certain diseases and developed correct diagnoses of them; proposed new concepts of hygiene; made use of anesthetics in surgery with newly innovated surgical tools; and introduced the science of dissection in anatomy. They furthered the scientific breeding of horses and cattle; found new ways of grafting to produce new types of flowers and fruits; introduced new concepts of irrigation, fertilization, and soil cultivation; and improved upon the science of navigation. In the area of chemistry, Muslim scholarship led to the discovery of such substances as potash, alcohol, nitrate of silver, nitric acid, sulfuric acid, and mercury chloride. It also developed to a high degree of perfection the arts of textiles, ceramics, and metallurgy.

The Islāmic renaissance and medieval periods

Major periods of Muslim education and learning. The renaissance of Islāmic culture and scholarship developed largely under the ʿAbbāsid administration in eastern Islām and under the later Umayyads in western Islām, mainly in Spain, between 800 and 1000. This latter period, the golden age of Islāmic scholarship, was largely a period of translation and interpretation of classical thoughts and their adaptation to Islāmic theology and philosophy. The period also witnessed the introduction and assimilation of Hellenistic, Persian, and Hindu mathematics, astronomy, algebra, trigonometry, and medicine into Muslim culture.

Whereas the 8th and 9th centuries, mainly between 750 and 900, were characterized by the introduction of classical learning and its refinement and adaptation to Islāmic culture, the 10th and 11th were centuries of interpretation, criticism, and further adaptation. There followed a period of modification and significant additions to classical culture through Muslim scholarship. Then, during the 12th and 13th centuries, most of the works of classical learning and the creative Muslim additions were translated from Arabic into Hebrew and Latin. The decline of Muslim scholarship coincided with the early phases of the European intellectual awakening that these translations were partly instrumental in bringing about.

Creative scholarship in Islām from the 10th to the 12th century included works by such scholars as Omar Khayyam, al-Bīrūnī, Fakhr ad-Dīn ar-Rāzī, Avicenna (Ibn Sīnā), aṭ-Ṭabarī, Avempace (Ibn Bājjah), and Averroës (Ibn Rushd).

Influence of Islāmic learning on the West. The translation into Latin of most Islāmic works during the 12th and 13th centuries had a great impact upon the European Renaissance. As Islām was declining in scholarship and Europe was absorbing the fruits of Islām's centuries of creative productivity, signs of Latin Christian awakening were evident throughout the European continent.

The 12th century was one of intensified traffic of Muslim learning into the Western world through many hundreds of translations of Muslim works, which helped Europe seize the initiative from Islām when political conditions in Islām brought about a decline in Muslim scholarship. By 1300, when all that was worthwhile in Muslim scientific, philosophical, and social learning had been transmitted to European schoolmen through Latin translations, European scholars stood once again on the solid ground of Hellenistic thought, enriched or modified through Muslim and Byzantine efforts. (M.K.N./J.S.Sz.)

The European Middle Ages

THE BACKGROUND OF EARLY CHRISTIAN EDUCATION

From the beginnings to the 4th century. At first Christianity found most of its adherents among the poor and illiterate, making little headway, as St. Paul observed (1 Corinthians 1:26), among the worldly-wise, the mighty, and those of high rank. But during the 2nd century AD and afterward it appealed more and more to the educated class and to leading citizens. These naturally wanted their children to have at least as good an education as they themselves had had, but the only schools available were the grammar and rhetoric schools, with their Greco-Roman, non-Christian culture. There were different opinions among Christian leaders about the right attitude to this dilemma that confronted all Christians who sought a good education for their children. The Greek Fathers, especially the Christian Platonists Clement of Alexandria and Origen, sought to prove that the Christian view of the universe was compatible with Greek thought and even regarded Christianity as the culmination of philosophy, to which the way must be sought through liberal studies. Without a liberal education the Christian could live a life of faith and obedience but could not expect to attain an intellectual understanding of the mysteries of the faith or expect to appreciate the significance of the Gospel as the meeting ground of Hellenism and Judaism. St. Augustine and St. Basil also tolerated the use of the secular schools by Christians, maintaining that literary and rhetorical culture is valuable so long as it is kept subservient to the Christian life. The Roman theologian Tertullian, on the other hand, was suspicious of pagan culture, but he admitted the necessity, though deploring it, of making use of the educational facilities available.

Views of the early Christian Fathers

In any event, most Christians who wanted their children to have a good education appear to have sent their children to the secular schools; this practice continued even after 313, when the emperor Constantine, who had been converted to Christianity, stopped the persecution of Christians and gave them the same rights as other citizens. Christians also set up catechetical schools for the religious instruction of adults who wished to be baptized. Of these schools the most famous was the one at Alexandria in Egypt, which had a succession of outstanding heads, including Clement and Origen. Under their scholarly guidance, it developed a much wider curriculum than was usual in catechetical schools, including the best in Greek science and philosophy in addition to Christian studies. Other schools modeled on that at Alexandria developed in some parts of the Middle East, notably in Syria, and continued for some time after the collapse of the empire in the west.

From the 5th to the 8th century. The gradual subjugation of the Western Empire by the barbarian invaders during the 5th century eventually entailed the breakup of the educational system that the Romans had developed over the centuries. The barbarians, however, did not destroy the empire; in fact, their entry was really in the form of vast migrations that swamped the existing and rapidly weakening Roman culture. The position of the emperor remained, the barbarians exercising local control through smaller kingdoms. Roman learning continued, and there were notable examples in the writings of Boethius, chiefly his *Consolation of Philosophy.* Boethius composed most of these studies while acting as director of civil administration under the Ostrogoths. Equally famous was his contemporary Cassiodorus (*c.* 490–*c.* 585) who, as a minister under the Ostrogoths, worked energetically at his vision of *civilitas,* a program of educating the public and developing a sound administrative structure. Thus, despite the political and social upheavals, the methods and program of ancient education survived into the 6th century in the new barbarian Mediterranean kingdoms; indeed, the barbarians were frequently impressed and attracted by things Roman. In Ostrogothic Italy (Milan, Ravenna, Rome) and in Vandal Africa (Carthage), the schools of the grammarians and rhetoricians survived for a time, and, even in those places where such schools soon disappeared, as in Gaul and in Spain, private teachers or parents maintained the tradition of classical culture until the 7th century. As in previous centuries, the culture bestowed was essentially literary and oratorical: grammar and rhetoric constituted the basis of the studies. The pupils read, reread, and commented on the classical authors and imitated them by composing certain kinds of exercises (*dictiones*) with the aim of achieving a perfect mastery of their style. In fact, however, the practice was desultory, and the results were mechanical and poor. Greek was ignored more and more, and attempts to revive Hellenic studies were limited to a dwindling number of scholars.

Christianity, meanwhile, was becoming more formally organized, and in the Latin-speaking western division of the empire, the Catholic church (as it was beginning to

be called, from the Greek *katholikos,* the "whole") had developed an administrative pattern, based upon that of the empire itself, for which learning was essential for the proper discharge of its duties. Schools began to be formed in the rudimentary cathedrals, although the main centres of learning from the 5th century to the time of Charlemagne in the 8th century were in the monasteries. The prototype of Western monasticism was the great monastery founded at Monte Cassino in 529 by Benedict of Nursia (*c.* 480–*c.* 547), probably on the model of Vivarium, the scholarly monastery established by Cassiodorus. The rule developed by Benedict to guide monastic life stimulated many other foundations, and one result was the rapid spread of Benedictine monasteries and the establishment of an order. The Benedictine monasteries became the chief centres of learning and the source of the many literate scribes needed for the civil administration.

Monastic and episcopal schools

The monastic schools, however, are no more significant in the history of education than the schools founded by bishops, usually in connection with a cathedral. These episcopal schools are sometimes looked upon as successors of the grammar schools of the Roman Empire. First specializing in the development of the clergy, they later admitted young lay people when the small Roman schools had disappeared. At the same time there were bishops who organized a kind of boarding school where the aspiring clergyman, living in a community, participated in duties of a monastic character and learned his clerical trade.

The influence of monasticism affected the content of instruction and the method of presenting it. Children were to be dutiful, as the Celtic and English monks Columban and Bede were to remark: "A child does not remain angry, he is not spiteful, does not contradict the professors, but receives with confidence what is taught him." In the case of the adolescent destined for a religious profession, the monastic lawgiver was severe. The teacher must know and teach the doctrine, reprimand the undisciplined, and adapt his method to the different temperaments of the young monks. The education of young girls destined for monastic life was similar: the mistress of the novices recommended prayer, manual work, and study.

Education of the laity

Between the 5th and 8th centuries the principles of education of the laity likewise evolved. The treatises on education, later called the "mirrors," pointed to the importance of the four moral virtues—prudence, courage, justice, and temperance. The *Institutionum disciplinae* of an anonymous Visigoth pedagogue expressed the desire that all young men "quench their thirst at the quadruple fountain of the virtues." In the 7th and 8th centuries the moral concepts of antiquity completely surrendered to religious principles. The Christian Bible was more and more considered as the only source of moral life, as the mirror in which men must learn to see themselves. A bishop addressing himself to a son of the Frankish king Dagobert (died 639) drew his examples from the books of the Old Testament. The mother of Didier of Cahors addressed to her son letters of edification on the fear of God, on the horror of vice, and on penitence.

The Christian education of children who were not aristocrats or future clergymen or monks was irregular. Whereas in antiquity catechetical instruction was organized especially for the adult laity, after the 5th century more and more children and then infants received baptism, and, once baptized, a child was not required to receive any particular religious education. His parents and godparents assisted him in learning the minimum, if anything at all. Only by attending church services and listening to sermons did the child acquire his religious culture.

The Irish and English revivals. During the 5th and 6th centuries there was a renaissance of learning in the remote land of Ireland, introduced there initially by the patron saints of Ireland—Patrick, Bridget, and Columba—who established schools at Armagh, Kildare, and Iona. They were followed by a number of other native scholars, who also founded colleges—the most famous and greatest university being the one at Clonmacnois, on the Shannon River near Athlone. To these and lesser schools flocked Anglo-Saxons, Gauls, Scots, and Teutons from Britain and the Continent. From about AD 600 to 850, Ireland itself sent scholars to the Continent to teach, found monasteries, and establish schools.

Although the very earliest Irish scholars may have aimed primarily at propagating the Christian faith, their successors soon began studying and teaching the Greek and Roman classics (but only in Latin versions), along with Christian theology. Eventually there were additions of mathematics, nature study, rhetoric, poetry, grammar, and astronomy, all studied, it seems, very largely through the medium of the Irish language.

England was next to experience the reawakening, and, though there were notable schools at such places as Canterbury and Winchester, it was in Northumbria that the schools flourished most. At the monasteries of Jarrow and Wearmouth and at the Cathedral School of York, some of the greatest of early medieval writers and schoolmasters appeared, including the Venerable Bede and Alcuin. The latter went to France in 780 to become master of Charlemagne's palace school.

THE CAROLINGIAN RENAISSANCE AND ITS AFTERMATH

The cultural revival under Charlemagne and his successors. Charlemagne (742/743–814) has been represented as the sponsor or even creator of medieval education, and the Carolingian renaissance has been represented as the renewal of Western culture. This renaissance, however, built on earlier episcopal and monastic developments; and, although Charlemagne did help to ensure the survival of scholarly traditions in a relatively bleak and rude age, there was nothing like the general advance in education that occurred later with the cultural awakening of the 11th and 12th centuries.

Educational reforms of Charlemagne

Learning, nonetheless, had no more ardent friend than Charlemagne, who came to the Frankish throne in 768 distressed to find extremely poor standards of Latin prevailing. He thus ordered that the clergy be educated severely, whether by persuasion or under compulsion. He recalled that, in order to interpret the Holy Scriptures, one must have a command of correct language and a fluent knowledge of Latin; he later commanded, "in each bishopric and in each monastery let the psalms, the notes, the chant, calculation and grammar be taught and carefully corrected books be available" (capitulary of AD 789). His promotion of ecclesiastical and educational reform bore fruit in a generation of churchmen whose morals and whose education were of a higher standard than before.

The possibility then arose of providing, for the brighter young clerics and perhaps also for a few laymen, a more advanced religious and academic training. It was perhaps to meet this modest need that a school grew up within the precincts of the emperor's palace at Aachen. In order to develop and staff other centres of culture and learning, Charlemagne imported considerable foreign talent. During the 8th century England had been the scene of some intellectual activity; thus, Alcuin, who had been the master of the school at York, and other English scholars were brought over to transplant to the Continent the studies and disciplines of the Anglo-Saxon schools. From Moorish Spain came Christian refugees who also contributed to this intellectual revival; disputations with the Muslims had forced them to develop a dialectic skill in which they now instructed Charlemagne's subjects. From Italy came grammarians and chroniclers, men such as Paul the Deacon; the more formalistic classical traditions in which they had been bred supplied the framework to discipline the effervescent brilliance of the Anglo-Saxons. Irish scholars also arrived. Thanks to these foreigners, who represented the areas where classical and Christian culture had been maintained in the 6th and 8th centuries, the court became a kind of "academy," to use Alcuin's term. There the emperor, his heirs, and his friends discussed various subjects—the existence or nonexistence of the underworld and of nothingness, the eclipse of the sun, the relationship of Father, Son, and Holy Spirit, and so on. Recognizing the importance of manuscripts in the cultural revival, Charlemagne formed a library (the catalog of which is still extant), had texts and books copied and recopied, and bade every school to maintain a scriptorium. Alcuin developed a school of calligraphy at Tours, and its new

script spread rapidly throughout the empire; this Carolingian minuscule was more legible and less wasteful of space than the uncial scripts hitherto employed.

Outside the court at Aachen were to be found here and there a few seats of culture, but not many. The archbishop of Lyon reorganized the schools of readers and choir leaders; Alcuin in Saint-Martin-de-Tours and Angilbert in Saint-Riquier organized monastic schools with relatively well-stocked libraries. It was necessary to wait for the second generation or even the third to witness the greatest brilliance of the Carolingian renewal. Under Charlemagne's son Louis the Pious and especially under his grandsons, the monastic schools reached their apogee in France north of the Loire, in Germany, and in Italy. The most famous were at Saint-Gall, Reichenau, Fulda, Bobbio, Saint-Denis, Saint-Martin-de-Tours, and Ferrières. Unfortunately, the breakup of the Carolingian empire, following local rebellions and the Viking invasions, ended the progress of the Carolingian renaissance.

Educational accomplishments of Alfred the Great

Influences of the Carolingian renaissance abroad. In England, at least in the kingdom of Wessex, King Alfred the Great stands out as another royal patron of learning, one who wanted to imitate the creativity of Charlemagne. When he came to the throne in 871, cultural standards had fallen to a low level, partly because of the turmoil of the Danish invasions. He was grieved to find so few who could understand Latin church services or translate a letter from Latin into English. To accomplish an improvement, he called upon monks from the Continent, particularly those of Saint-Bertin. Moreover, he attracted to his court certain English clergy and young sons of nobles. Since the latter did not know Latin, he had translated into Wessex English some works of Pope Gregory the Great, Boethius, the theologian and historian Paulus Orosius, Venerable Bede, St. Augustine, and others. He himself translated Boethius' *Consolation of Philosophy,* Gregory the Great's *Pastoral Care,* and Bede's *Ecclesiastical History of the English People.* This promotion of learning was continued by Alfred's successors and spread elsewhere in England; and in the reformed monasteries at Canterbury, York, and Winchester, the young monks renewed the study of religious and secular sciences. Among the master scholars of the late 10th century was the Benedictine monk Aelfric, perhaps the greatest prose writer of Anglo-Saxon times. In order to facilitate the learning of Latin for young monks, Aelfric composed a grammar, glossary, and colloquy, containing a Latin grammar described in Anglo-Saxon, a glossary in which master and pupil could find a methodically classified Latin vocabulary (names of birds, fish, plants, and so forth), and a manual of conversation, inspired by the bilingual manuals of antiquity.

Among the other Saxons, those of the Continent who presided over the destinies of Germany, there were also significant gatherings of masters and students at selected monasteries, such as Corvey and Gandersheim. In any case, wherever teaching became important in the 10th century, it concentrated largely on grammar and the works of the classical authors. Thus when Gerbert of Aurillac, after a course of instruction in Catalonia, came to teach dialectic and the arts of the quadrivium (geometry, arithmetic, harmonics, and astronomy) at Reims, he aroused astonishment and admiration. His renown helped in his later election as Pope Sylvester II. The first half of the 11th century contained the first glimmerings of a rediscovered dialectic. A new stage in the history of teaching was beginning.

Religious and military emphases of general education

Education of the laity in the 9th and 10th centuries. The clergy who dominated society thought it necessary to give laymen some directives about life comparable to those offered in monastic rules and thus issued what were called *miroirs* ("mirrors"), setting forth the duties of a good sovereign and exalting the Christian struggle. Already the image of the courtly and Christian knight was beginning to take shape. It was not a question of governing a state well but, rather, of governing oneself. The layman must struggle against vice and practice virtue; he must emphasize his religious heritage. Alcuin became indignant when he heard it said that the reading of the Gospel was the duty of the clergy and not that of the layman. Huoda, wife of Bernard, duke of Septimania, addressed a manual to her 16-year-old son, stressing the reading and praying that a young man should do. In the libraries of the laity, the volumes of the Old and New Testaments took first place, along with prayer books, a kind of breviary designed for day-to-day use.

If a minority of aristocrats could receive a suitable moral and religious education, the masses remained illiterate and preferred a military apprenticeship to study. "He who has remained in school up to twelve years without mounting a horse is no longer good for anything but the priesthood," wrote a German poet. Writers of hagiographic texts were fond of contrasting the mother of the future saint, anxious to give education to her son, and the father, who wanted to harden his son at an early age to the chase or to war. The Carolingian tradition, however, was not totally forgotten by princes and others in high places. In Germany, Otto I and his successors, who wished to re-create the Carolingian empire, encouraged studies at the court: Wipo, the preceptor of Henry III, set out a program of education for the laity in his *Proverbia.* Rediscovering the ancient moralists, chiefly Cicero and Seneca, he praised moderation as opposed to warlike brutality or even the ascetic strength of the monks. The same tendency is found in other writings.

THE MEDIEVAL RENAISSANCE

The era that has been called the "renaissance of the 12th century" corresponds to a rediscovery of studies originating in the 11th century in a West in the process of transformation. The church cast off the tutelage of lay power, and there was general acceptance of the authority of the church in matters of belief, conduct, and education; the papacy took over the direction of Christianity and organized the Crusades to the East; the monarchies regrouped the political and economic forces of feudal society; the cities were reanimated and were organized into communes; the merchants traced out the great European trade routes and, before long, the Mediterranean ones. Soon contact with the East, by trade and in the Crusades, and with the highly cultivated Moors in Spain further stimulated intellectual life. Arabic renderings of some of the works of Aristotle, together with commentaries, were translated into Latin, exercising a profound influence on the trend of culture. It was inevitable that the world of education would take on a new appearance.

Decline of monastic studies

Changes in the schools and philosophies. *Monastic schools.* In the first place, the monastic reformers made the decision to close their schools to those who did not intend to enter upon a cloistered life. According to their idea of solitude and sanctity, recalling the words of St. Jerome, "the monk was not made to teach but to mortify himself." Divine works were to be the only object of study and meditation, and Pierre de Celle asserted that "divine science ought to mould rather than question, to nourish conscience rather than knowledge."

The scholarly monks completed their studies before being admitted to the monastery—the age of entrance in Benedictine houses, for instance, being fixed at 15 years at Cîteaux and 20 years at Cluny. If there were admitted a few oblates (who were laymen living in monasteries under modified rules), they were given an ascetic and moral education and were taught to read the Holy Writ and, what was still more desirable, to "relish" it. In the Carthusian monastery the four steps of required spiritual exercise were reading, meditation, prayer, and contemplation. Thus, there existed a monastic culture, but there were no truly monastic studies such as those that had existed in the 9th and 10th centuries. The rich libraries of the monasteries served only a few scholarly abbots, while the monks searched for God through prayer and asceticism.

Urban schools. In the cities, on the contrary, the schools offered to all the clergy who so desired the means of satisfying their intellectual appetite. More and more of them attended these schools, for the studies were a good means of social advancement or material profit. The development of royal and municipal administrations offered the clergy new occupations. Hence the success of the schools for notaries and the schools of law and rhetoric.

These schools were organized under the protection of the collegiate churches and the cathedrals. The schools for secular subjects were directed by an archdeacon, chancellor, cantor, or cleric who had received the title of *scholasticus, caput scholae,* or *magister scholarum* and who was assisted by one or more auxiliary masters. The success of the urban schools was such that it was necessary, in the middle of the 12th century, to define the teaching function. Only those could teach who were provided with the *licencia docendi* conferred by the bishop or, more often, by the *scholasticus.* Those who were licensed taught within the limits of the city or the diocese, whose clerical leaders supervised this monopoly and intervened if a cleric set himself up as master without having the right. The popes were sufficiently concerned about licensing that the Lateran Council of 1179 gave this institution universal application.

New curricula and philosophies. The pupils who attended these urban schools learned in them their future occupation as clerics; they learned Latin, learned to sing the various offices, and studied Holy Writ. The more gifted ones extended their studies further and applied for admission to the liberal arts (the trivium, made up of grammar, rhetoric, and logic; and the quadrivium, including geometry, arithmetic, harmonics, and astronomy) and, upon completion of the liberal arts, to philosophy. Philosophy had four branches: theoretical, practical, logical, and mechanical. The theoretical was divided into theology, physics, and mathematics; the practical consisted of morals or ethics (personal, economic, political). The logical, which concerned discourse, consisted of the three arts of the trivium. Finally, the mechanical included the work of processing wool, of navigation, of agriculture, of medicine, and so on. This was an ambitious humanistic program. In fact, the students became specialized in the study of one art or another according to their tastes or the presence of a renowned master, such as Guillaume de Champeaux at Paris and St. Victor for rhetoric and theology; Peter Abelard at Paris for dialectic and theology; Bernard de Chartres for grammar; William of Conches at Chartres for grammar, ethics, and medicine; and Thierry de Chartres for rhetoric. In particular, teachers of the "literary" arts, grammar and rhetoric, always had great success in a period of enthusiasm for the ancient authors. It may be noted that Bernard de Chartres organized his literary teaching in this fashion: grammatical explanations (*declinatio*), studies of authors, and each morning the correction of the exercises given the day before.

Study of the liberal arts and philosophy

The third art of the trivium, logic (or dialectics), was nevertheless a strong competitor of the other two, grammar and rhetoric. Since the 11th century, Aristotle's *Posterior Analytics,* which had been translated centuries earlier by Boethius, had developed the taste for reasoning, and, by the time that Abelard arrived in Paris around 1100, interest in dialectics was flourishing. The written words of the Scriptures and of the Fathers of the Church were to be subjected to the scrutiny of human reason; a healthy skepticism was to be the stepping-stone to knowledge, aided by an understanding of critical logic. While dialectic reigned in Paris, the masters at Chartres offered a study of the whole of the quadrivium. This interest in the sciences, which had been manifest at Chartres since the early 11th century, had been favoured by the stimulus of Greco-Arabic translations. The works of Euclid, Ptolemy, Hippocrates, Galen, and other Hellenic and Hellenistic scholars, as preserved in the Arabic manuscripts, were translated in southern Italy, Sicily, and Spain and were gradually transmitted northward. The scientific revival allowed the Chartrians to Christianize Greek cosmology, to explain Genesis according to physics, and to rediscover nature. Another revival was that of law. The conflicts in the second half of the 12th century between the church and the lay powers encouraged on both sides a new activity in the juridical field. The princes found in the Corpus Juris Civilis, the 6th-century Roman code of the emperor Justinian, the means of legitimizing their politics, and the papacy likewise used Roman sources to promote its claims.

Thomist philosophy. In the long view, the greatest educational and philosophical influence of the age was St. Thomas Aquinas, who in the 13th century made a monumental attempt to reconcile the two great streams of the Western tradition. In his teaching at the University of Paris and in his writings—particularly the *Summa theologiae* and the *Summa contra gentiles*—Aquinas tried to synthesize reason and faith, philosophy and theology, university and monastery, activity and contemplation. In his writings, however, faith and theology ultimately took precedence over reason and philosophy because the former were presumed to give access to truths that were not available through rational inquiry. Hence, Aquinas started with assumptions based on divine revelation and went on to a philosophical explication of man and nature. The model of the educated man that emerged from this process was the Scholastic, a man whose rational intelligence had been vigorously disciplined for the pursuit of moral excellence and whose highest happiness was found in contemplation of the Christian God.

The Scholastic model

The Scholastic model greatly affected the development of Western education, especially in fostering the notion of intellectual discipline. Aquinas' theological-philosophical doctrine was a powerful intellectual force throughout the West, being officially adopted by the Dominican order (of which Aquinas was a member) in the 13th century and by the Jesuits in the 17th century. Known as Thomism, this doctrine came to constitute the basis of official Roman Catholic theology from 1879. Although Aquinas made an important place in his hierarchy of values for the practical uses of reason, later Thomists were often more exclusively intellectual in their educational emphasis.

The development of the universities. The Middle Ages were thus beset by a multiplicity of ideas, both homegrown and imported from abroad. The multiplicity of students and masters, their rivalries, and the conflicts in which they opposed the religious and civil authorities obliged the world of education to reorganize. To understand the reorganization, one must review the various stages of development in the coming together of students and masters. The first stage, already alluded to, occurred when the bishop or some other authority began to accord to other masters permission to open schools other than the episcopal school in the neighbourhood of his church. A further stage was reached when a license to teach, the *jus ubique docendi*—granted only after a formal examination—empowered a master to carry on his vocation at any similar centre. A further development came when it began to be recognized that, without a license from pope, emperor, or king, no school could be formed possessing the right of conferring degrees, which originally meant nothing more than licenses to teach.

Students and teachers, as *clerici* ("clerks," or members of the clergy), enjoyed certain privileges and immunities, but, as the numbers traveling to renowned schools increased, they needed additional protection. In 1158 Emperor Frederick I Barbarossa of the Holy Roman Empire granted them privileges such as protection against unjust arrest, trial before their peers, and permission to "dwell in security." These privileges were subsequently extended and included protection against extortion in financial dealings and the *cessatio,* or the right to strike, discontinue lectures, and even to secede to protest against grievances or interference with established rights.

In the north of Europe licenses to teach were granted by the chancellor, *scholasticus,* or some other officer of a cathedral church; in the south it is probable that the guilds of masters (when these came to be formed) were at first free to grant their own licenses, without any ecclesiastical or other supervision. Gradually, however, toward the end of the 12th century, a few great schools, from the excellence of their teaching, came to assume more than local importance. In practice, a doctor of Paris or Bologna would be allowed to teach anywhere; and those great schools began to be known as studia generalia; that is, places resorted to by scholars from all parts. Eventually the term came to have a more definite and technical significance. The emperor Frederick II in 1225 set the example of attempting to confer upon his new school at Naples, by an authoritative bull, the prestige that the earlier *studia* had acquired by reputation and general consent. Pope Gregory IX did the

The studia generalia and the *universitas*

same for Toulouse in 1229, and he added to its original privileges in 1233 a bull by which anyone who had been admitted to the doctorate or mastership in that university should have the right to teach anywhere without further examination. Other studia generalia were subsequently founded by papal or imperial bulls, and in 1292 even the oldest universities, Paris and Bologna, found it desirable to obtain similar bulls from Pope Nicholas IV. From this time the notion began to prevail that the essence of the studium generale was the privilege of conferring a universally valid teaching license and that no new *studium* could acquire that position without a papal or imperial bull. There were, however, a few studia generalia (such as Oxford) the position of which was too well established to be questioned, even though they had never obtained such a bull; these were held to be studia generalia by repute. A few Spanish universities founded by royal charter were held to be studia generalia for the kingdom.

The word *universitas* originally applied only to the scholastic guild (or guilds)—that is, the corporation of students and masters—within the *studium,* and it was always modified, as *universitas magistrorum,* or *universitas scholarium,* or *universitas magistrorum et scholarium.* In the course of time, however, probably toward the latter part of the 14th century, the term began to be used by itself, with the exclusive meaning of a self-regulating community of teachers and scholars whose corporate existence had been recognized and sanctioned by civil or ecclesiastical authority.

The Italian universities. The earliest *studia* arose out of efforts to provide instruction beyond the range of the cathedral and monastic schools for the education of priests and monks. Salerno, the first great *studium,* became known as a school of medicine as early as the 9th century, and, under the teaching of Constantine the African (died 1087), its fame spread throughout Europe. In 1231 it was licensed by Frederick II as the only school of medicine in the kingdom of Naples. It remained a medical school only.

The Bologna prototype of universities

The great revival of legal studies that took place at Bologna about the year 1000 had been preceded by a corresponding activity at Pavia and Ravenna. In Bologna a certain Pepo was lecturing on parts of the Corpus Juris Civilis about the year 1076. The secular character of this new study and its close connection with the claims and prerogatives of the Western emperor aroused papal suspicion, and for a time Bologna and its students were regarded by the church with distrust. The students found their first real protector in the emperor Frederick I Barbarossa. The immunities and privileges he conferred eventually extended to all the other universities of Italy.

The first university of Bologna was not constituted until the close of the 11th century—the "universities" there being student guilds, formed to obtain by combination that protection and those rights that they could not claim as citizens. As the number of students increased, the number of *universitates,* or societies of scholars, increased, each representing the national origin of its members (France, England, Provence, Spain, Italy). These confederations were presided over by a common head, the *rector scholarium,* and the different nations were represented by their *consiliarii,* a deliberative assembly with which the rector habitually took counsel. The practice at Bologna was adopted as other studia generalia arose.

The students at Bologna were mostly of mature years. Because civil law and canon law were, at first, the only branches of study offered, the class they attracted was often composed of lawyers already filling office in some department of the church or state—archdeacons, heads of schools, canons of cathedrals, and like functionaries. About 1200 the two faculties of medicine and philosophy were formed. The former was developed by a succession of able teachers, among whom Thaddeus Alderottus was especially eminent. The faculty of arts, down to the 14th century, scarcely attained equal eminence.

At Bologna the term college long had a different meaning from the ordinary modern one. The masters formed themselves into collegia (that is, organizations), chiefly for the conferment of degrees. Places of residence for students existed at Bologna at a very early date, but it was not until the 14th century that they possessed any organization; the humble *domus,* as it was termed, was at first designed solely for necessitous students who were not natives of Bologna; a separate house, with a fund for the maintenance of a specified number of scholars, was all that was originally contemplated.

From the 13th to the 15th century a number of universities in Italy originated from migrations of students; others were established by papal or other charters. Almost all the schools taught civil or canon law or both. Of these institutions the most important were Padua, Piacenza, Pavia, Rome, Perugia, Pisa, Florence, Siena, and Turin.

The Paris prototype of universities

The French universities. The history of the University of Paris well illustrates the fact that the universities arose in response to new needs. The schools out of which the university arose were those attached to the Cathedral of Notre-Dame de Paris on the Île de la Cité and presided over by its chancellor. Although, in the second decade of the 13th century, some masters placed themselves under the jurisdiction of the abbot of the monastery of Sainte-Geneviève on the Left Bank of the Seine, it was around the bestowal of the license by the chancellor of Notre-Dame that the university grew. It is in this license that the whole significance of the master of arts degree was contained; for admission to that degree was the receiving of the chancellor's permission to "incept"; and by "inception" was implied the master's formal entrance upon the functions of a duly licensed teacher and his recognition as such by his brethren in the profession. The stage of bachelordom had been one of apprenticeship for the mastership; and his emancipation from this state was symbolized by the placing of the magisterial cap (biretta) upon his head. The new master gave a formal inaugural lecture, and he was then welcomed into the society of his professional brethren with set speeches and took his seat in his master's chair.

Some time between 1150 and 1170 the University of Paris came formally into being. Its first written statutes were not, however, compiled until about 1208, and it was not until long after that date that it possessed a "rector." Its earliest recognition as a legal corporation belongs to about the year 1211, when a brief of Innocent III empowered it to elect a proctor to be its representative at the papal court. With papal support Paris became the great transalpine centre of orthodox theological teaching. Successive pontiffs, down to the Great Schism of 1378, cultivated friendly relations with the university and systematically discouraged the formation of theological faculties at other centres. In 1231 Gregory IX, in the bull *Parens scientiarum* ("Mother of Learning"), gave full recognition to the right of the several faculties to regulate and modify the constitution of the university. The fully developed university was divided into four faculties: three superior, those of theology, canon law, and medicine; and one inferior, that of arts, which was divided into four student confederations, or nations (French, Picard, Norman, and English), which included both professors and scholars from the respective countries. The head of each faculty was the dean; of each nation, the proctor. The rector, in the first instance head of the faculty of arts, eventually became the head of the collective university.

After the close of the Middle Ages, Paris came to be virtually reduced to a federation of colleges, though at Paris the colleges were less independent of university authority than was often the case elsewhere. Other major French universities of the Middle Ages were Montpellier, Toulouse, Orléans, Angers, Avignon, Cahors, Grenoble, Orange, and Perpignan.

The English offshoots of Paris: Oxford and Cambridge

The English universities. The University of Paris became the model for French universities north of the Loire and for those of central Europe and England; Oxford would appear to have been the earliest. Certain schools, opened early in the 12th century within the precincts of the dissolved nunnery of St. Frideswide and of Oseney Abbey, are supposed to have been the nucleus around which it grew. But the beginning may have been a migration of English students from Paris about 1167 or 1168. Immediately after 1168, allusions to Oxford as a *studium* and a studium generale begin to multiply. In the 13th century, mention first occurs of university "chests," which

were benefactions designed for the assistance of poor students. Halls, or places of licensed residence for students, also began to be established. Against periodic vicissitudes such as student dispersions and plagues, the foundation of colleges proved the most effective remedy. The earliest colleges were University College, founded in 1249, Balliol College, founded about 1263, and Merton College, founded in 1264.

The University of Cambridge, although it came into existence somewhat later than Oxford, may reasonably be held to have had its origin in the same century. In 1112 the canons of St. Giles crossed the River Cam and took up their residence in the new priory in Barnwell, and their work of instruction acquired additional importance. In 1209 a body of students migrated there from Oxford. Then about 1224 the Franciscans established themselves in the town and, somewhat less than half a century later, were followed by the Dominicans. At both the English universities, as at Paris, the mendicants and other religious orders were admitted to degrees, a privilege that, until the year 1337, was extended to them at no other university. Their interest in and influence at these three centres were consequently proportionately great.

In 1231 and 1233 royal and papal letters afford satisfactory proof that the University of Cambridge was already an organized body, with a chancellor at its head.

Although both Oxford and Cambridge were modeled on Paris, their higher faculties never developed the same distinct organization; and, while the two proctors at Cambridge originally represented north and south, the nations are scarcely to be discerned. An important step was made, however, in 1276, when an ordinance was passed requiring that everyone who claimed to be recognized as a scholar should have a fixed master within 15 days after his entry into the university. The traditional constitution of the English universities was, in its origin, an imitation of the Parisian, modified by the absence of the cathedral chancellor. But the feature that most served to give permanence and cohesion to the entire community at Cambridge was, as at Oxford, the institution of colleges. The earliest of these was Peterhouse, in 1284. All the early colleges were expressly designed for the benefit of the secular clergy.

Universities elsewhere in Europe. From the 13th to the 15th centuries, studia generalia or universities proliferated in central and northern Europe and were usually modeled on the University of Paris. Although the earliest was Prague, which existed as a *studium* in the 13th century and was chartered by Pope Clement VI in 1348, perhaps no medieval university achieved a more rapid and permanent success than Heidelberg. The University of Heidelberg, the oldest in the German realm, received its charter in 1386 from Pope Urban VI as a studium generale and contained all the recognized faculties—theology, canon law, medicine, and the arts, as well as civil law. In the subsequent 100 years, universities were founded at Cologne, Erfurt, Leipzig, Rostock, Freiburg, Tübingen, Ofen (Budapest), Basel, Uppsala, and Copenhagen.

Higher education in Spain

Spain was also an important scene of developments in higher education. Valladolid received its charter in 1346 and attained great celebrity after it obtained the rank of studium generale and a *universitas theologiae* by a decree of Pope Martin V in 1418. Salamanca was founded in 1243 by Ferdinand III of Castile with faculties of arts, medicine, and jurisprudence, to which theology was added through the efforts of Martin V. The College of St. Bartholomew, the earliest founded at Salamanca, was noted for its ancient library and valuable collection of manuscripts. Other important early Spanish and Portuguese schools were Seville, Alcalá, and Lisbon.

General characteristics of medieval universities. Generally speaking, the medieval universities were conservative. Alexander Hegius and Rudolf Agricola carried on their work as reformers at places such as Deventer, in the Netherlands, remote from university influences. A considerable amount of mental activity went on in the universities; but it was mostly of the kind that, while giving rise to endless controversy, turned upon questions in connection with which the implied postulates and the terminology employed rendered all scientific investigation hopeless. At almost every university the realists and nominalists represented two great parties occupied with an internecine struggle (see EPISTEMOLOGY).

In Italian universities such controversies were considered endless and their effects pernicious. It was resolved, accordingly, to expel logic and allow its place to be filled by rhetoric, thereby effecting that important revolution in academic studies that constituted a new era in university learning and largely helped to pave the way for the Renaissance. The professorial body in the great Italian universities attained an almost unrivaled reputation throughout Europe. For each subject of importance there were always two, and sometimes three, rival chairs. While other universities became sectarian and local, those of Italy continued to be universal, and foreigners of all nations could be found among the professors.

The material life of the students was difficult. In order to aid the poorest, some colleges founded by clerical or lay benefactors offered board and lodging to a number of foundationers. Courses, too, could occasionally be difficult. The courses in theology were particularly long—eight years at the minimum (one could not be a teacher of theology in Paris before the age of 35). Many students preferred the more rapid and more lucrative paths of law and medicine. Others led the life of perpetual students, of vagabond clerics, disputatious goliards, the objects of repeated but ineffectual condemnation.

The methods of teaching are particularly well known in the case of Paris. The university year was divided into two terms: from St. Remi (October 1) to Lent and from Easter to St. Pierre (June 29). The courses consisted of lectures (*collatio*) but more often of explications of texts (*lectio*). There were also discussions and question periods. Examinations were given at the end of each term. The student could receive three degrees: the *determinatio,* or baccalaureate, gave him the right to teach under the supervision of a master; the *licencia docendi* was literally the "license to teach" and could be obtained at 21 years of age; then there was the doctorate, which marked his entrance into mastership and which involved a public examination.

Lay education and the lower schools. The founding of universities was naturally accompanied by a corresponding increase in schools of various kinds. In most parts of western Europe, there were soon grammar schools of some type available for boys. Not only were there grammar schools at cathedrals and collegiate churches, but many others were founded in connection with chantries and craft and merchant guilds and a few in connection with hospitals. It has been estimated, for example, that, toward the close of the Middle Ages, there were in England and Wales, for a population of about 2.5 million, approximately 400 grammar schools, although the number of their enrollments was generally quite small.

Early grammar schools

In fulfillment of its responsibility for education, the church from the 11th century onward made the establishment of an effective education system a central feature of ecclesiastical policy. During the papacy of Gregory VII (1073–85), all bishops had been asked to see that the art of grammar was taught in their churches, and a Lateran Council in 1215 decreed that grammar-school masters should be appointed not only in the cathedral church but also in others that could afford it. Solicitude at the centre for the advancement of education did not, however, result in centralized administration. It was the duty of bishops to carry out approved policy, but it was left to them to administer it, and they in turn allowed schools a large measure of autonomy. Such freedom as medieval schools enjoyed was, however, always subject to the absolute authority of the church, and the right to teach, as earlier noted, was restricted to those who held a bishop's license. This device was used to ensure that all teachers were loyal to the doctrines of the church.

Knowledge of the teaching provided in the grammar schools at this period is too slight to justify an attempt at a description. No doubt the curriculum varied, but religion was all-important, with Latin as a written and spoken language the other major element in the timetable. There might have been instruction in reading and writing in the vernacular, but, in addition to the grammar schools,

there were writing and song schools and other schools of an elementary type. Elementary teaching was given in many churches and priests' houses, and children who did not receive formal scholastic instruction were given oral teaching by parish priests in the doctrines and duties of the faith. The evidence of accounts, bills, inventories, and the like suggests that there was some careful teaching of writing and of an arithmetic that covered the practical calculations required in ordinary life. Literacy, however, was limited by the lack of printed materials; until the 15th century (when typesetting developed) books were laboriously cut page by page on blocks (hence they were known as block books) and consequently were rare and expensive. From the mid-15th century on, literacy increased as typeset books became more widely available.

Educational provision for girls in medieval society was much more restricted. Wealthy families made some provision in the home, but the emphasis was primarily on piety and secondarily on skills of household management, along with artistic "accomplishments." Neither girls nor boys of the lowest social ranks—peasants or unskilled urban dwellers—were likely to be literate. Nor were girls of the artisan classes until the 16th century, when female teaching congregations such as the Ursulines founded by Angela Merici began to appear. There were, however, provisions for boys of the artisan class to receive sufficient vernacular schooling to enable them to be apprenticed to various trades under the auspices of the guilds.

Training for feudalism

There was an entirely different training for boys of high rank, and this created a cultural cleavage. Instead of attending the grammar school and proceeding to a university, these boys served as pages and then as squires in the halls and castles of the nobility, there receiving prolonged instruction in chivalry. The training was designed to fit the noble youth to become a worthy knight, a just and prudent master, and a sensible manager of an estate. Much of this knowledge was gained from daily experience in the household, but, in addition, the page received direct instruction in reading and writing, courtly pastimes such as chess and playing the lute, singing and making verses, the rules and usages of courtesy, and the knightly conception of duty. As a squire he practiced more assiduously the knightly exercises of war and peace and acquired useful experience in leadership by managing large and small bodies of men. But this was a type of education that could flourish only in a feudal society; and, though some of its ideals survived, it was outmoded when feudalism was undermined by the growth of national feeling.

(P.R./J.Bo.)

Education in Asian civilizations: c. 700 to the eve of Western influence

INDIA

During its medieval period, India was ruled by dynasties of Muslim culture and religion. Muslims from Arabia first appeared in the country in the 8th century, but the foundation of their rule was laid much later by Muḥammad Ghūri, who established his power at Delhi in 1192. The original Muslim rule was replaced successively by that of the Muslim Pashtuns and Mughals.

The foundations of Muslim education. Muslim educational institutions were of two types—a *maktab,* or elementary school, and a madrasah, or institution of higher learning. The content of education imparted in these schools was not the same throughout the country. It was, however, necessary for every Muslim boy at least to attend a *maktab* and to learn the necessary portions of the Qur'ān required for daily prayers. The curriculum in the madrasah comprised Ḥadīth (the study of Muslim traditions), jurisprudence, literature, logic and philosophy, and prosody. Later on, the scope of the curriculum was widened, and such subjects as history, economics, mathematics, astronomy, and even medicine and agriculture were added. Generally, all the subjects were not taught in every institution. Selected madrasahs imparted postgraduate instruction, and a number of towns—Āgra, Badaun, Bīdar, Gulbarga, Delhi, Jaunpur, and a few others—developed into university centres to which students flocked for study under renowned scholars. The sultans and amirs of Delhi and the Muslim rulers and nobles in the provinces also extended patronage to Persian scholars who came from other parts of Asia under the pressure of Mongol inroads. Delhi vied with Baghdad and Córdoba as an important centre of Islāmic culture. Indian languages also received some attention. The Muslim rulers of Bengal, for example, engaged scholars to translate the Hindu classics, the *Rāmāyaṇa* and the *Mahābhārata,* into Bengali.

Developments under the Pathans

Under the Pathan Lodis, a dynasty of Afghan foreigners (1451–1526), the education of the Hindus was not only neglected but was often adversely affected in newly conquered territories. The rulers generally tolerated Sanskrit and vernacular schools already in existence but did not help the existing ones with money or build new ones. At early stages, the *maktab*s and madrasahs were attended by Muslims only. Later, when Hindus were allowed into high administrative positions, Hindu children began to receive Persian education in Muslim schools.

The Mughal period. The credit for organizing education on a systematic basis goes to Akbar (lived 1542–1605), a contemporary of Queen Elizabeth I of England and undoubtedly the greatest of Mughal emperors. He treated all his subjects alike and opened a large number of schools and colleges for Muslims as well as for Hindus throughout his empire. He also introduced a few curricular changes, based on students' individual needs and the practical necessities of life. The scope of the curriculum was so widened as to enable every student to receive education according to his religion and views of life. The adoption of Persian as the court language gave further encouragement to the Hindus and the Muslims to study Persian.

Akbar's policy was continued by his successors Jahāngīr and Shāh Jahān. But his great-grandson Aurangzeb (1618–1707) changed his policy with regard to the education of the Hindus. In April 1669, for instance, he ordered the provincial governors to destroy Hindu schools and temples within their jurisdiction; and, at the same time, he supported Muslim education with a certain religious fanaticism. After his death, the glory of the Mughal empire began gradually to vanish, and the whole country was overrun by warlords.

During the Mughal period, girls received their education at home or in the house of some teacher living in close proximity. There were special arrangements for the education of the ladies of the royal household, and some of the princesses were distinguished scholars. Vocational education was imparted through a system of apprenticeship either in the house of *ustād*s (teachers) or in *kārkhānah*s (manufacturing centres).

Muslim rulers of India were also great patrons of literature and gave considerable impetus to its development. Akbar ordered various Hindu classics and histories translated into Persian. In addition, a number of Greek and Arabic works were translated into Persian. Literary activities did not entirely cease even in the troubled days of later rulers. Men of letters were patronized by such emperors as Bahādur Shāh and Muḥammad Shāh and by various regional officials and landlords.

Such is the history of Muslim education in India. It resembles ancient Indian education to a great extent: instruction was free; the relation between the teachers and the taught was cordial; there were great centres of learning; the monitorial system was used; and people were preoccupied with theology and the conduct of life. There were, however, several distinctive features of Muslim education.

Educational accomplishments of the Muslim period

First, education was democratized. As in mosques, so in a *maktab* or madrasah, all were equal, and the principle was established that the poor should also be educated. Second, Muslim rule influenced the system of elementary education of the Hindus, which had to accommodate itself to changed circumstances by adopting a new method of teaching and by using textbooks full of Persian terms and references to Muslim usages. Third, the Muslim period brought in many cultural influences from abroad. The courses of studies were both widened and brought under a humanistic influence. Finally, Muslim rule produced a cross-cultural influence in the country through the establishment of an educational system in which Hindus and

Muslims could study side by side and in which there would be compulsory education in Persian, cultivation of Sanskrit and Hindi, and translation of great classics of literature into different languages. Ultimately, it led to the development of a common medium of expression, Urdu.

Education in the Muslim era was not a concerted and planned activity but a voluntary and spontaneous growth. There was no separate administration of education, and state aid was sporadic and unsteady. Education was supported by charitable endowments and by lavish provision for the students in a madrasah or in a monastery.

The Muslim system, however, proved ultimately harmful. In the early stages genuine love of learning attracted students to the cultural centres, but later on "the bees that flocked there were preeminently drones." The whole system became stagnant and stereotyped as soon as cultural communication was cut off from the outside world because of political disturbances and internecine wars. The Indian teachers were reduced to dependence on their own resources, and a hardening tradition that became increasingly unreceptive to new ideas reduced the whole process to mere routine. (S.N.M.)

CHINA

The T'ang dynasty (AD 618–907). The T'ang was one of China's greatest dynasties, marked by military power, political stability, economic prosperity, and advance in art, literature, and education. It was an age in which Buddhist scholarship won recognition and respect for its originality and high intellectual quality and in which China superseded India as the land from which Buddhism was to spread to other countries in East Asia.

The T'ang was known for its literature and art and has been called the golden age of Chinese poetry. There were thousands of poets of note who left a cultural legacy unsurpassed in subsequent periods and even in other lands. Prose writers also flourished, as did artists whose paintings reflected the influences of Buddhism and Taoism.

One of the greatest gifts of China to the world was the invention of printing. Block printing was invented in the 8th century and movable type in the 11th century. The first book printed from blocks was a Buddhist sutra, or set of precepts, in 868. Printing met the demand created by the increase in the output of literature and by the regularized civil service examination system. It also met the popular demand for Buddhist and Taoist prayers and charms. One historian (Kenneth Scott Latourette) noted that "as late as the close of the eighteenth century the [Chinese] Empire possibly contained more printed books than all the rest of the world put together."

Education in the T'ang dynasty was under the dominant influence of Confucianism, notwithstanding the fact that Buddhism and Taoism both received imperial favours. A national academic examination system was firmly established, and officials were selected on the basis of civil service examinations. But Confucianism did not dominate to the extent of excluding other schools of thought and scholarship. Renowned scholars were known to spurn public office because they were not satisfied with a narrow interpretation of Confucianism. Artists and poets were, in general, rebellious against traditional Confucianism.

An emperor in the 5th century ordered the establishment of a "School of Occult Studies" along with the more commonly accepted schools of Confucian learning. It was devoted to the study of Buddhism and Taoism and occult subjects that transcended the practical affairs of government and society. Such schools were often carried on by the private effort of scholars who served as tutors for interested followers.

School structure in the T'ang era

The schools of T'ang were well organized and systematized. There were schools under the central government, others under local management, and private schools of different kinds. Public schools were maintained in each prefecture, district, town, and village. In the capital were "colleges" of mathematics, law, and calligraphy, as well as those for classical study. There was also a medical school.

Semiprivate schools formed by famous scholars gave lectures and tutelage to students numbering in the hundreds. Students from Korea and Japan came to study in China and took back the lunar calendar and the Buddhist sects, as well as the examination system and the Confucian theories of government and social life. Chinese culture also penetrated Indochina.

The examination system was at this time given the form that remained essentially unchanged until the 20th century. Examinations were held on different levels, and for each a corresponding academic degree was specified. Interestingly, there was provision for three degrees, not unlike the bachelor's, master's, and doctor's degrees of modern times. The first degree was the *hsiu ts'ai* ("cultivated talent"), the second the *ming ching* ("understanding the classics"), and the third the *chin-shih* ("advanced scholar"). The name of the second degree was in later periods changed to *chü jen* ("recommended man"). An academy of scholars later known as the Hanlin Academy was established for select scholars whom the emperor could call upon for advice and expert opinion on various subjects. Membership in this institution became the highest honour that could be conferred upon those who passed the *chin-shih* degree with distinction. To be appointed a Hanlin scholar was to be recognized as one of the top scholars of the land. Among the services that they rendered were the administration and supervision of examinations and the explanation of difficult texts in literature, classics, and philosophy.

Examinations were given for students of medicine and for military degrees. The study of medicine included acupuncture and massage, as well as the treatment of general diseases of the body and those of eye, ear, throat, and teeth.

The Sung (960–1279). The Sung was another dynasty of cultural brilliance. Landscape painting approached perfection, and cultural achievement was stimulated by the invention of movable type (first made of earthenware, then of wood and metal). This advance from the older method of block printing led to the multiplication of books; the printing of a complete set of the classics was a boon to literary studies in schools.

The rulers of Sung were receptive to new ideas and innovative policies. The outstanding innovator of the dynasty was Wang An-shih, prime minister from 1068 to 1076. He introduced a comprehensive program of reform that included important changes in education; more emphasis was subsequently placed on the study of current problems and political economy.

Educational controversy in the Sung era

Wang's reforms met with opposition from conservatives. The controversy was only a phase of a deeper and more far-reaching intellectual debate that made the philosophical contributions of the Sung scholars as significant as those of the Hundred Schools in the Chou dynasty over a millennium earlier. Confucianism and the dominant mode of Chinese thinking had been subject to the challenge of ideas from legalism, Taoism, and Buddhism, and, despite the resistance of conservatives, the traditional views had to be modified. Outstanding Confucian scholars of conservative bent argued vigorously with aggressive proponents of new concepts of man, of knowledge, and of the universe. The result was Neo-Confucianism, or what some prefer to call rational philosophy. The most eminent Neo-Confucianist was Chu Hsi, a Confucian scholar who had studied Taoism and Buddhism. His genius lay in his ability to synthesize ideas from a fresh point of view. Sung scholars distinguished themselves in other fields, too. Ssu-ma Kuang's *Tzu-chih t'ung-chien* ("Comprehensive Mirror for Aid in Government") was a history of China from the 5th century BC to the 10th century AD. The result of 20 years of painstaking research, it consisted of 1,000 chapters prepared under imperial direction. A volume on architecture was produced that is still used today as a basic reference work, and a treatise on botany contained the most ancient record of varieties of citrus fruits then known in China. No less worthy of mention is an encyclopaedia titled *T'ai-ping yü lan.*

The general pattern of the school system remained essentially the same, with provision for lower schools, higher schools, and technical schools, but there was a broadening of the curriculum. A noteworthy development was the rise of a semiprivate institution known as the *shu-yüan,* or academy. With financial support coming from both state grants and private contributions, these academies were

Academies

managed by noted scholars of the day and attracted many students and lecturers. Often located in mountain retreats or in the woods, they symbolized the influence of Taoism and Buddhism and a desire to pursue quiet study far away from possible government interference.

The Mongol period (1206–1368). The Mongols were ferocious fighters but inept administrators. Distrustful of the Chinese, they enlisted the services of many nationalities and employed non-Chinese aliens. To facilitate the employment of these aliens, the civil service examinations were suspended for a number of years. Later, when a modified form of examinations was in effect, there were special examinations for Mongol candidates to make sure of their admission into high offices.

The Mongols despised the Chinese and placed many limitations on them. Consequently, an aftermath of Mongol rule was a strong antiforeign reaction on the part of the Chinese, accompanied by an overanxious desire to preserve the Chinese heritage.

Despite the setback in Chinese culture under Mongol rule, the period was not devoid of positive cultural development. The increase in foreign contacts as a result of travel to and from China brought new ideas and new knowledge of other lands and other peoples. Mathematics and medicine were further influenced by new ideas from abroad. The classics were translated into the Mongol language, and the Mongol language was taught in schools.

Private schools and the academies of the Sung dynasty became more popular. As a result of a decrease in opportunities for government appointment, scholars withdrew into the provinces for study and tutoring. Relieved of the pressure of preparing for the examinations, they applied their talents to the less formal but more popular arts and literary forms, including the drama and the novel. Instead of the classical form, they used the vernacular, or the spoken, language. The significance of this development was not evident until the 20th century, when a "literary revolution" popularized the vernacular tongue.

The Ming period (1368–1644). The Ming dynasty restored Chinese rule. Ming was famous for its ceramics and architecture. There were excellent painters, too, but they were at best the disciples of the T'ang and Sung masters. The outstanding intellectual contribution of the period was the novel, whose development was spurred by increases in literacy and in the demand for reading materials. Ming novels are today recognized as masterpieces of popular vernacular literature. Also of note was the compilation of *Pen-ts'ao kang-mu* ("Great Pharmacopoeia"), a valuable volume on herbs and medicine that was the fruit of 26 years of labour.

The *Yung-lo* encyclopaedia

Of considerable scholarly and educational importance was the *Yung-lo ta-tien* encyclopaedia, which marked a high point in the Chinese encyclopaedic movement. It was a gigantic work resulting from the painstaking efforts of 2,000 scholars over a period of five years. It ran into more than 11,000 volumes, too costly to print, and only two extra copies were made.

The examination system remained basically the same. In the early period of the dynasty, the schools were systematized and regularized. In the latter part of the dynasty, however, the increasing importance of the examination system relegated the schools to a secondary position. The decline of the state-supported schools stimulated the further growth of private education.

The Manchu period (1644–1911/12). Except for two capable emperors, who ruled for a span of 135 years at the beginning, the Manchu dynasty was weak and undistinguished. Under Emperors K'ang-hsi and Ch'ien-lung, learning flourished, but there was little originality. The alien Manchu rulers concentrated on the preservation of what seemed best for stability and the maintenance of the status quo. They wanted new editions of classical and literary works, not creative contributions to scholarship.

Distrust of the Chinese by the Manchus and a feeling of insecurity caused the conquerors to erect barriers between themselves and the Chinese. The discriminatory policy was expressed in the administration of the examinations. To assure the appointment of Manchus to government posts, equal quotas were set aside for the Manchus and the Chinese, although the former constituted only about 3 percent of the population. The Chinese thus faced the keenest competition in the examinations, and those who passed tended to be brilliant intellects, whereas the Manchus could be assured of success without great effort.

Schools were encouraged and regulated during the early period of the dynasty. The public school system consisted of schools for nobles, national schools, and provincial schools. Separate schools were maintained for the Manchus, and, for their benefit, Chinese books were translated into the Manchu language. Village and charitable schools were supported by public funds, but they were neglected in later years; so that, by the end of the dynasty, private schools and tutoring had overshadowed them.

At the threshold of the modern era, China had sunk into political weakness and intellectual stagnation. The creativity and originality that had brightened previous periods of history were now absent. Examinations dominated the educational scene, and the content of the examinations was largely literary and classical. Taoism and Buddhism had lost their intellectual vigour, and Confucianism became the unchallenged model of scholarship.

Much could be said for the Chinese examination system at its best. It was instrumental in establishing an intellectual aristocracy whereby the nation could be sure of a cultural unity by entrusting government to scholars reared in a common tradition, nurtured in a common cultural heritage, and dedicated to common ideals of political and social life. It established a tradition of government by civilians and by scholars. It made the scholars the most highly esteemed people of the land. The examinations provided an open road to fame and position. Chinese society was not without classes, but there was a high degree of social mobility, and education provided the opportunity for raising one's position and status. There were no rigid prerequisites and no age limits for taking the examinations. Selection was rigorous, but the examinations were, on the whole, administered with fairness. The names of the candidates did not appear on the examination papers, and the candidates were not permitted to have any outside contacts while writing them.

Influence of the examination system

Nevertheless, the system had serious drawbacks. The content of the examinations became more and more limited in scope. The Confucianist classics constituted the core, and a narrow and rigid interpretation prevailed. In early times, Chinese education was broad and liberal, but, by the 19th century, art, music, and science had been dropped on the wayside; even arithmetic was not accorded the same importance as reading and writing. Modern science and technology were completely neglected.

After alien rule by the Mongols the Chinese were obsessed with restoring their heritage; they avoided deviating from established forms and views. This conservatism was accentuated under Manchu rule and resulted in sterility and stagnation. The creativity and original spirit of classical education was lost. The narrow curriculum was far removed from the pressing problems and changing needs of the 19th century. (T.H.C.)

JAPAN

The ancient period to the 12th century. The Japanese nation seems to have formed a unified ancient state in the 4th century AD. Society at that time was composed of shizoku, or clans, each of which served the *chōtei* ("the imperial court") with its specialized skill or vocation. People sustained themselves by engaging in agriculture, hunting, and fishing, and the chief problem of education was how to convey the knowledge of these activities and provide instruction in the skills useful for these occupations.

The influence of the civilizations of China and India had a profound effect on both the spiritual life and the education of the Japanese. Toward the 6th century the assimilation of Chinese civilization became more and more rapid, particularly as a result of the spread of Confucianism. Buddhism was also an important intellectual and spiritual influence. Originating in India and then spreading to China, Buddhism was transmitted to Japan through the Korean peninsula in the mid-6th century.

Influence of Chinese and Indian civilization

A monarchic state system with an emperor as its head

was established following a coup d'état in 645. The subsequent Taika (Great Reform) era saw the beginning of many new institutions, most of which were primarily imitations of institutions of the T'ang dynasty of China. In the field of education, a *daigakuryō,* or college house, was established in the capital, and *kokugaku,* or provincial schools, were built in the provinces. Their chief aim was to train government officials. The early curriculum was almost identical to that of the T'ang dynasty of China but by the 8th and 9th centuries had been modified considerably to meet internal conditions, particularly as regards the educational needs of the nobility.

Through the Nara and the Heian eras (8th to 12th century), the nobility (*kuge*) constituted the ruling class, and learning and culture were the concern primarily of the *kuge* and the Buddhist monks. The *kuge* lived an artistic life, so that the emphasis of education came to be placed on poetry, music, and calligraphy. Teaching in the *daigakuryō* gradually shifted in emphasis from Confucianism to literature, since the *kuge* set a higher value on artistic refinement than on more spiritual endeavours. Apart from the *daigakuryō,* other institutions were established in which families of influential clans lodged and developed their intellectual lives.

The feudal period (1192–1867). *Education of the warriors.* Toward the mid-12th century political power passed from the nobility to the *buke,* or warrior, class. The ensuing feudal period in Japan dates from the year 1192 (the establishment of the Kamakura shogunate) to 1867 (the decline of the Tokugawa shogunate).

The warrior's way of life was quite unlike that of the nobility, and the aims and content of education in the warrior's society inevitably differed. The warrior had constantly to practice military arts, hardening his body and training his will. Education was based on military training, and a culture characteristic of warriors began to flourish. Some emphasis, though, was placed on spiritual instruction. The warrior society, founded on firm master–servant relations and centring on the philosophy of Japanese family structure, set the highest value on family reputation and on genealogies. Furthermore, because the military arts proved insufficient to enable warriors to grasp political power and thereby maintain their ruling position, there arose a philosophy of *bumbu-kembi,* which asserted the desirability of being proficient in both literary and military arts. Thus, the children of warriors attended temples and rigorously trained their minds and wills. Reading and writing were the main subjects.

Temples were the centres of culture and learning and can be said to have been equivalent to universities, in that they provided a meeting place for scholars and students. Education in the temple, originally aimed at instructing novitiates, gradually changed its character, eventually providing education for children not destined to be monks. Thus, the temples functioned as institutions of primary education.

Education in the Tokugawa era. In 1603 a shogunate was established by a warrior, Tokugawa Ieyasu, in the city of Edo (present Tokyo). The period thence to the year 1867, the Tokugawa, or Edo, era, constitutes the later feudal period in Japan. This era, though also dominated by warriors, differed from former ones in that internal disturbances finally ended and long-enduring peace ensued. There emerged a merchant class that developed a flourishing commoner's culture. Schools for commoners thus were established.

Temple schools of the Tokugawa period

Representative of such schools were the *terakoya* (temple schools), deriving from the earlier education in the temple. As time passed, some *terakoya* used parts of private homes as classrooms. Designed to be one of the private schools, or *shijuku,* the *terakoya* developed rapidly in the latter half of the Tokugawa era, flourishing in most towns and villages. Toward the end of the era they assumed the characteristics of the modern primary school, with emphasis on reading, writing, and arithmetic. Other *shijuku,* emphasizing Chinese, Dutch, and national studies, as well as practical arts, contributed to the diversification of learning and permitted students with different class and geographic backgrounds to pursue learning under the guidance of the same teacher. Their curricula were free from official control.

The shogunate established schools to promote Confucianism, which provided the moral training for upper-class samurai that was essential for maintaining the ideology of the feudal regime. *Han,* or feudal domains, following the same policy, built *hankō,* or domain schools, in their castle towns for the education of their own retainers.

The officially run schools for the samurai were at the apex of the educational system in the Tokugawa era. The Confucian Academy, which was known as the Shōheikō and was administered directly by the shogunate, became a model for *hankō* throughout Japan. The *hankō* gradually spread after about 1750, so that by the end of the era they numbered over 200.

The curriculum in the *hankō* consisted chiefly of *kangaku* (the study of books written in Chinese) and, above all, of Confucianism. Classics of Confucianism, historical works, and anthologies of Chinese poems were used as textbooks. Brush writing, *kokugaku* (study of thought originating in Japan), and medicine were also included. Later, in the last days of the shogunate, *yōgaku,* or Western learning, including Western medicine, was added in several institutions.

Both *hankō* for samurai and *terakoya* for commoners were the typical schools after the middle of the Tokugawa era. Also to be found, however, were *gōgaku,* or provincial schools, for samurai as well as commoners. They were founded at places of strategic importance by the feudal domain.

Role played by private schools

The various *shijuku* became centres of interaction among students from different domains when such close contact among residents of different areas was prohibited. They served as centres of learning and dialogue for many of those who later constituted the political leadership responsible for the Meiji Restoration of 1868.

Effect of early Western contacts. The Europeans who first arrived in Japan were the Portuguese, in 1543. In 1549 the Jesuit Francis Xavier visited Japan, and, for the first time, the propagation of Christianity began. Many missionaries began to arrive, Christian schools were built, and European civilization was actively introduced.

In 1633 the shogunate, in apprehension of further Christian infiltration of Japan, banned foreign travel and prohibited the return of overseas Japanese. Further, in 1639, the shogunate banned visits by Europeans. This was the so-called *sakoku,* or period of national isolation. From that time on Christianity was strictly forbidden, and international trade was conducted with only the Chinese and the Dutch. Because contact with Europeans was restricted to the Dutch, Western studies developed as *rangaku,* or learning through the Dutch language.

It is noteworthy that the Tokugawa period laid the foundation of modern Japanese learning. As a result of the development of *hankō* and *terakoya,* Japanese culture and education had developed to such an extent that Japan was able to absorb Western influences and attain modernization at a remarkably rapid pace after the Meiji Restoration. (A.Na./N.S.)

European Renaissance and Reformation

THE CHANNELS OF DEVELOPMENT IN RENAISSANCE EDUCATION

The Muslim influence. Western civilization was profoundly influenced by the rapid rise and expansion of Islām from the 7th until the 15th century. By 732, 100 years after the death of Muḥammad, Islām had expanded from western Asia throughout all of northern Africa, across the straits of Gibraltar into Spain, and into France, reaching Tours, halfway from the Pyrenees to Paris. Muslim Spain rapidly became one of the most advanced civilizations of the period, where much of the learning of the past—Oriental, Greek, and Roman—was preserved and further developed. In particular, Greek and Latin scholarship was collected in great libraries in the splendid cities of Córdoba, Seville, Granada, and Toledo, which became major centres of advanced scholarship, especially in the practical arts of medicine and architecture.

Inevitably, scholarship in the adjacent Frankish, and subsequent French, kingdom was influenced, leading to a revitalization of western Christian scholarship, which had long been dormant as a result of the barbarian migrations. The doctrines of Aristotle, which had been assiduously cultivated by the Muslims, were especially influential for their emphasis on the role of reason in human affairs and on the importance of the study of humankind in the present, as distinct from the earlier Christian preoccupation with the cultivation of faith as essential for the future life. Thus, Muslim learning helped to usher in the new phase in education known as humanism, which first took definite form in the 12th century.

Humanism

The secular influence. The word humanism comes from *studia humanitatis* ("studies of humanity"). Toward the end of the Middle Ages there was a renewed interest in those studies that stressed the importance of man, his faculties, affairs, worldly aspirations, and well-being. The primacy of theology and otherworldliness was over; the *reductio artium ad theologiam* (freely, "reducing everything to theological argument") was rejected since it no longer expressed the reality of the new situation that was developing in Europe, particularly in Italy. Society had been profoundly transformed, commerce had expanded, and life in the cities had evolved. Economic and political power, previously in the hands of the ecclesiastical hierarchy and the feudal lords, was beginning to be taken over by the city burghers. Use of the vernacular languages was becoming widespread. The new society needed another kind of education and different educational structures; the burghers required new instruments with which to express themselves and found the old medieval universities inadequate.

The educational institutions of humanism had their origin in the schools set up in the free cities in the late 13th and the 14th centuries—schools designed to answer to the needs of the new urban population that was beginning to have greater economic importance in society. The pedagogical thought of the humanists took these transformations of society into account and worked out new theories that often went back to the classical Greek and Latin traditions; it was not, however, a servile imitation of the pedagogical thought and institutions of the classical world.

The Renaissance of the classical world and the educational movements it gave rise to were variously expressed in different parts of Europe and at various times from the 14th to the 17th century; there was a connecting thread, but there were also many differences. What the citizens of the Florentine republic needed was different from what was required by princes in the Renaissance courts of Italy or in other parts of Europe. Common to both, however, was the rejection of the medieval tradition that did not belong in the new society they were creating. Yet the search for a new methodology and a new relation with the ancient world was bitterly opposed by the traditionalists, who did not want renewal that would bring about a profound transformation of society; and, in fact, the educational revolution did not completely abolish existing traditions. The humanists, for example, were not concerned with extending education to the masses but turned their attention to the sons of princes and rich burghers.

The humanists had the important and original conception that education was neither completed at school nor limited to the years of one's youth but that it was a continuous process making use of varied instruments: companionship, games, and pleasure were part of education. Rather than suggesting new themes, they wanted to discover the method by which the ancient texts should be studied. For them knowledge of the classical languages meant the possibility of penetrating the thought of the past; grammar and rhetoric were being transformed into philological studies not for the sake of pedantic research but in order to acquire a new historical and critical consciousness. They reconstructed the past in order better to understand themselves and their own time.

THE HUMANISTIC TRADITION IN ITALY

Early influences. One of the most influential of early humanists was Manuel Chrysoloras, who came to Florence from Constantinople in 1396. He introduced the study of Greek and, among other things, translated Plato's *Republic* into Latin, which were important steps in the development of the humanistic movement.

Inspired by the ancient Athenian schools, the Platonic Academy established in Florence in the second half of the 15th century became a centre of learning and diffusion of Christian Platonism, a philosophy that conceived of all forms as the creative thoughts of God and that inspired considerable artistic innovation and creativity. Marsilio Ficino and Pico della Mirandola were two of the most original of the scholars who taught there. Florence was the first city to have such a centre, but Rome and Naples soon had similar academies, and Padua and Venice also became centres of culture.

A famous early humanist and professor of rhetoric at Padua was Pietro Paolo Vergerio (1370–1444). He wrote the first significant exclusively pedagogical treatise, *De ingenuis moribus et liberalibus studiis* ("On the Manners of a Gentleman and on Liberal Studies"), which, though not presenting any new techniques, did set out the fundamental principles by which education should be guided. He gave pedagogical expression to the ideal of harmony, or equilibrium, found in all aspects of humanism, and underlined the importance of the education of the body as well as of the spirit. The liberal arts were emphasized ("liberal" because of the liberation they reputedly brought); the program outlined by Vergerio focused upon eloquence, history, and philosophy but also included the sciences (mathematics, astronomy, and natural science) as well as medicine, law, metaphysics, and theology. The later subjects were not studied in depth; humanism was by its nature against encyclopaedism, but it brought out the relations between the disciplines and enabled students to know many subjects before they decided in which to specialize. Learning was not to be exclusively from books, and emphasis was placed on the advantages of preparing for social life by study and discussion in common. Vergerio felt that education should not be used as a means of entering the lucrative professions; medicine and law, especially, were looked on with suspicion if one's aim in studying them was merely that of gaining material advantages.

Emphasis on the liberal arts

Emergence of the new gymnasium. As a result of the renewed emphasis on Greek studies, early in the 15th century a definite sequence of institutions emerged, with the gymnasium as the principal school for young boys, preparatory to further liberal studies in the major nonuniversity institution of higher learning, the academy. Both terms, gymnasium and academy, were classical revivals, but their programs were markedly different from those of ancient Greece. The gymnasiums appeared in ducal courts; they were created for the liberal education of privileged boys and as the first stage of the *studia humanitatis.* Outstanding among these early gymnasiums were the school conducted by Gasparino da Barzizza in Padua from 1408 to 1421, considered a model for later institutions, and more particularly the gymnasium of Guarino Veronese (1374–1460) and that of his contemporary Vittorino da Feltre (1378–1446).

Guarino had first established a school in 1415 in Venice, where he was joined by Vittorino. He subsequently moved to Ferrara where, from 1429 to 1436, he assumed responsibility for the humanist education of the young son of Nicolò d'Este, the lord of Ferrara. Guarino wrote no treatises, but something may be learned about his work and methods from his large correspondence and from *De ordine docendi et studendi* (1485; "On the Order for Teaching and Studying"), written by his son Battista. Guarino organized his students' courses into three stages: the elementary level, at which reading and pronunciation were primarily taught, followed by the grammatical level, and finally the highest level, concentrating on rhetoric. The education given in his schools was perhaps the best example of the humanistic ideals, since it underlined the importance of literary studies together with a harmonious development of body and spirit, to the exclusion of any utilitarian purpose.

Humanist de-emphasis of "useful" education

Vittorino was a disciple of both Barzizza and Guarino. He conducted boarding schools at Padua and Venice and,

most importantly, from 1423 to 1446 one at Mantua, where he had been invited by the reigning lord, Gianfrancesco Gonzaga. This last school, known as La Giocosa (literally, "The Jocose, or Joyful"), soon became famous. At La Giocosa only those who had both talent and a modest disposition were accepted; wealth was neither necessary nor sufficient to gain admission; in fact, the school was one of the few efforts made during this period to extend education to a wider public. The program of study at La Giocosa was perhaps closer to the medieval tradition than that of the other boarding schools, but, in any case, the spirit was different. Studies were stimulating; mathematics was taught pleasantly—Vittorino going back to very ancient traditions of practicing mathematics with games. After having studied the seven arts of the trivium (grammar, rhetoric, and logic) and the quadrivium (geometry, arithmetic, harmonics, and astronomy), students completed the cycle by a study of philosophy and then, having mastered this discipline, could go on to higher studies leading to such professions as medicine, law, philosophy, and theology. Italian was completely ignored at Vittorino's school; all instruction was given in Latin, the study of which, together with Greek, reached a high level of excellence. Great importance was given to recreation and physical education; his concern for the health of his students did not come to an end with the scholastic year, for during the summers, when the cities became unhealthy, he would arrange for his students to go to Lake Garda or to the hills outside Verona.

Vittorino's educational philosophy was inspired by a profound religious faith and moral integrity, which contrasted with the general relaxation of standards within the church itself; but, if he was severe with himself, he was very open and tolerant with his pupils. The school continued only for a while after his death because, more than in the case of the other schools, La Giocosa was identified with the personality of the founder.

Nonscholastic traditions. Leon Battista Alberti, one of the most intelligent and original architects of the 15th century, also dedicated a treatise, *Della famiglia* (1435–44; "On the Family"), to methods of education. Alberti felt that the natural place for education was the home and not scholastic institutions. The language in which he wrote was Italian, education being in his view so important in social life that he felt that discussion of it should not be limited to scholars. He stressed the importance of the father in the educational process.

Baldassare Castiglione expressed the transition of humanism from the city to the Renaissance court. He himself was in the service of some of the most splendid princes, the Gonzagas at Mantua and the Montefeltros at Urbino. Just as in the 15th century the humanists had been concerned with the education of the city burgher, so in the 16th century they turned their attention to the education of the prince and of those who surrounded him. *Il cortegiano* ("The Courtier") was published in 1528, and within a few years it had been translated into Latin and all the major European languages. The courtier was to be the faithful collaborator of the prince. He had to be beautiful, strong, and agile; he had to know how to fight, play, dance, and make love. But this was not all, since great importance was also attached to the study of the classics and the practice of poetry and oratory; the courtier had to be able to write in rhyme and in prose and have perfect command of the vernacular, which was becoming important in political affairs; but above all he had to have skill at arms.

Education of the courtier

The courtier described by Castiglione, though in the service of necessarily devious princes, had to know how to keep his dignity and his virtue. Castiglione's moral standards, reflecting the spiritual climate at Urbino, completely disappeared, however, in Giovanni della Casa's work, *Galateo* (1551–54), in which considerations of etiquette were placed above all others; the values of humanism no longer existed, and all that was left was ceremonial.

THE HUMANISTIC TRADITION OF NORTHERN AND WESTERN EUROPE

The economic and social conditions behind the intellectual and cultural revolution of humanism in Italy were also present, though in different forms, in other parts of Europe. In some states, chiefly England, France, and Spain, humanism and educational reforms developed around the courts, where political power was being concentrated; in others, such as the Netherlands, they were brought about by the city burghers, whose power, both economic and political, was increasing. The educational reforms that the humanists brought about in northern and western Europe developed slowly, but on the whole they were lasting, since they affected a greater number of people than was the case in Italy, where they tended to be restricted to a narrow circle of families. There were close relations between Italian and other European educational humanists, as there were among English, Dutch, French, and German humanists, and, thus, national differences were not so significant.

Courtly and bourgeois forms of humanism

Dutch humanism. In the Netherlands the ground for educational reform had already been prepared in the 14th century by the Brethren of the Common Life, a group founded by Gerhard Groote to bring together laymen and religious men. Although their work was not originally in the field of education, education started when they set up hostels for students and exercised some moral direction over these students; this work was extended, and the Brethren eventually set up schools, first at Deventer, then in other cities. Some of the most important humanists of the Netherlands and Germany attended their schools—among others, Erasmus.

The school at Deventer came to have great prestige under Alexander Hegius, rector from 1465 to 1498 and author of a polemic treatise, *De utilitate Graeci* ("On the Usefulness of Greek"), underlining the importance of studying Greek, and of *De scientia* ("On Knowledge") and *De moribus* ("On Manners"). Hegius had great talent as an organizer and succeeded not only in attracting some of the best scholars of the time but also in giving the school an efficient structure that became a model for many schools in the north.

Desiderius Erasmus was a great scholar and educator, and his influence was felt all over Europe. His strong personality earned him the respect and sympathy of humanists who saw in him, as in few others, the symbol of their ideals and values. Unfortunately, his proposals for reform and greater tolerance were not always accepted in the tortured Europe of the 16th century.

Erasmus was a prolific writer, and part of his work was concerned with education: *De ratione studii* (1511; "On the Right Method of Study"), *De civilitate morum puerilium* (1526; "On the Politeness of Children's Manners"), *Ciceronianus* (1528), *De pueris statim ac liberaliter instituendis* (1529; "On the Liberal Education of Boys from the Beginning"). His educational program was original in many ways but in no sense democratic. The masses could not partake in higher education, since their aim was that of gaining skill in an occupation. He felt that religious instruction should be made available to all but that classical literary studies—the most important of all studies—were for a minority.

The classical and elitist tradition in humanism

Study of ancient languages and intelligent comprehension of texts formed the basis of Erasmus' system of education; he took a stand against the formalism and dogmatism that were already creeping into the humanist movement. Erasmus was in favour of acquiring a good general liberal arts education until the age of 18, being convinced that this would be a preparation for any form of further study. His great love for the classical languages, however, made him neglect the vernacular; he was not interested in local traditions; and he attributed very little importance to science, which he did not think necessary for a cultured man. He was against instruction being imposed without the participation of the student. His optimism about the nature of man and the possibilities of molding him made Erasmus feel that, if adequately educated, any man could learn any discipline. He further sought renewal of the schools and better training for teachers, which he felt should be a public obligation, certainly no less important than military defense. Many of Erasmus' themes were elaborated a century later by John Amos Comenius and form the basis of modern education, in particular the effort to understand the child psychologically and to consider education

as a process that starts before the school experience and continues beyond it.

Juan Luis Vives. Strongly influenced by Erasmus was Juan Luis Vives, who, though of Spanish origin, spent his life in various parts of Europe—Paris, Louvain, Oxford, London, Bruges. His most significant writings were *De institutione foeminae Christianae* (1523; "On the Education of a Christian Woman"), *De ratione studii puerilis* ("On the Right Method of Instruction for Children"), *De subventione pauperum* (1526; "On Aid for the Poor"), and *De tradendis disciplinis* (1531; "On the Subjects of Study").

The new social and utilitarian traditions

Not only was his vision of the organic unity of pedagogy new, but he was the first of the humanists to emphasize the importance of popular education. He felt that it was the responsibility of the city to provide instruction for the poor and that the craft and merchant guilds had an important contribution to make to education. Unlike other humanists, moreover, he did not despise the utilitarian aspects of education but on the contrary suggested that his pupils should visit shops and workshops and go out into the country to learn something of real life.

Just as he felt that education should not be limited to a single social class, so he felt that there should be no exclusion of women, though perhaps they required a different kind of education because of their different functions in life.

Vives worked out a plan to take account of both educational structures and teacher training. In emphasizing the social function of education, he was against schools being run for profit and believed that teachers should be prepared not only in their specific fields but also in psychology so as to understand the child. He also suggested that teachers should meet four times a year to examine together the intellectual capacities of each one of their pupils so that suitable programs of study could be arranged for them. Vives considered that, in teaching, games had psychological value. He favoured use of the vernacular for the first stage of education; but, as a humanist, he had a passion for Latin and felt that there was no substitute for Latin as a universal language. Classical studies were to be completed by investigation of the modern world, in particular its geography, the horizons having been greatly enlarged by recent discoveries. Vives' method was an inductive one, based not on metaphysical theories but on experiment and exercise.

The early English humanists. At the end of the 15th century there was a flowering in England of both humanistic studies and educational institutions, enabling a rapid transition from the medieval tradition to the Renaissance. The English humanists prepared excellent texts for studying the classical languages, and they started a new type of grammar school, long to be a model. Most important were John Colet and Thomas More. Thomas Linacre, author of *De emendata structura Latini sermonis libri sex* (1524; "Six Books on the Flawless Structure of the Latin Language"), should also be remembered, as well as William Lily, author of a Latin syntax, *Absolutissimus de octo orationis partium constructione libellus* (1515; "Comprehensive Study of the Construction of the Eight Parts of Speech"), and director of St. Paul's School in London from 1512 to 1522.

The English grammar schools

Colet has an important place in English education. As dean of St. Paul's Cathedral he founded St. Paul's School, thus favouring the introduction of humanism in England and the transformation of the old ecclesiastical medieval schools. He had traveled a great deal in France and Italy and wanted to bring to his country the humanistic culture that had so fascinated him. In 1510 he started a "grammar school," open to about 150 scholars who had an aptitude for study and had completed elementary school. Colet's personality and energy made his school a lively centre of English humanism.

More was both a distinguished humanist and a statesman. He was interested in pedagogy, to which he dedicated part of his work *Utopia* (1516). In his Utopia, More saw the connection between educational, social, and political problems and the influence that society therefore has on education. English humanists such as More were engaged in a bitter battle because medieval tradition was deeply rooted; they were fierce opponents of a group called the Trojans, who opposed the Greek language and all that the new instruction of that language represented.

EDUCATION IN THE REFORMATION AND COUNTER-REFORMATION

New political and social systems developed in those European countries that, for various reasons and at different times, broke away from the Roman Catholic church in the 16th century. The religious reforms brought about by Martin Luther, John Calvin, Huldrych Zwingli, and the ruling family of England were both cause and effect of these transformations. Characteristic of all these countries was the importance of the state in the organization of the educational system.

The Reformation and European humanism influenced one another. There were analogies between the flowering of the classical world in the European courts and the reawakening of religious interests; there were similarities in the critical position adopted toward Aristotelianism and in the interest shown toward the study of classical languages, such as Greek and Hebrew. The presuppositions behind the two movements—humanism and Reformation—were different, however, and sooner or later a clash was inevitable. The most spectacular of these clashes was between Erasmus and Luther, despite the fact that for a long time they had respected each other. It was important for Erasmus and for the humanists to encourage the development of a world of writers and artists who, free from material preoccupations, could devote their time to literary and artistic pursuits. For the Reformers the situation was different: they did not aim to educate a small minority; unlike Erasmus, Luther had to keep the masses in mind, for they had contributed to the success of the religious reforms.

Humanism and the Reformation

Luther and the German Reformation. Luther specifically wished his humble social origins to be considered a title of nobility. He wanted to create educational institutions that would be open to the sons of peasants and miners, though this did not mean giving them political representation. (The German princes were glad to promote the Reformation on condition that it would not diminish but would, on the contrary, increase their political power.) Luther realized that an educational system open to the masses would have to be public and financed by citizens' councils. His educational programs are set out in *An die Radsherrn aller Stedte deütsches Lands: Das sie christliche Schulen affrichten und hallten sollen* (1524; "Letter to the Mayors and Aldermen of All the Cities in Behalf of Christian Schools"), in *Dass man Kinder zur Schulen halten solle* (1530; "Discourse on the Duty of Sending Children to School"), and in various letters to German princes.

Although Luther advocated the study of classical languages, he believed that the primary purpose of such an education, in marked distinction to the aims of the humanists, was to promote piety through the reading of the Scriptures in their pure form. "Neglect of education," Luther wrote in a letter to Jacob Strauss in 1524, "will bring the greatest ruin to the Gospel." Accordingly, Luther argued that education must be extended to all children, girls as well as boys, and not simply to a leisured minority as in Renaissance Italy. Even those children who had to work for their parents in trade or in the fields should be enabled, if only for a few hours a day, to attend local, city-maintained schools in order to promote their reading skills and hence piety. Out of the Lutheran argument emerged a new educational concept, the *pietas litterata:* literacy to promote piety.

Lutheran emphasis on public schools for all

On the premise that a new class of cultivated men must be developed to substitute for the dispossessed monks and priests, new schools, whose upkeep was the responsibility of the princes and the cities, were soon organized along the lines suggested by Luther. In 1543 Maurice of Saxony founded three schools open to the public, supported by estates from the dissolved monasteries. It was more difficult to set up the city schools, for which there was no tradition. In towns and villages of northern Germany Johannes Bugenhagen (1485–1558) set up the earliest schools to teach religion and reading and writing in German, but it was

not until 1559 that the public ordinances of Württemberg made explicit reference to German schools in the villages. This example was shortly followed in Saxony.

Whereas Luther combined his interest in education with his work as a religious reformer and politician, another Reformer, Philipp Melanchthon (1497–1560), concentrated almost entirely on education, creating a new educational system and, in particular, setting up a secondary-school system. He taught for many years at the University of Wittenberg, which became one of the centres of theological studies in Reformation Germany; and his experience there enabled him to reorganize the old universities and set up new ones, such as Marburg, Königsberg, and Jena. His ideas about secondary education were put into practice in the schools he founded at Eisleben. Scholastic work was divided into three stages, access to each successive stage depending on the ability of the student to master the previous course work; this was a new concept (foretelling the later "grading system"), unknown in the traditional scholastic system. He was convinced that too many subjects should not be imposed on the student. He felt that Latin was important but not German, Greek, or Hebrew, as had been taught in the humanistic schools; such variety, he felt, was exhausting and possibly harmful. This opened the door to a new type of formalism, however, a danger that in other spheres the educational reformers had tried to fight.

Growing formalism in German schools

The work of Johannes Sturm (1507–89) illustrates this danger. He founded a grammar school in Strassburg (now Strasbourg, Fr.) that became a model for German schools. Sturm believed that methods of instruction in elementary schools and, to some degree, in secondary schools should be different from those in the institutes of higher education. Not much autonomy was to be allowed the child, who started learning Latin at the age of six by memorizing. Sturm's love of Latin was even greater than that of his friend Erasmus, who never wanted it to become a mechanical exercise. As a consequence, German was neglected, as was physical instruction, and too much importance was given to form and expression for its own sake.

The English Reformation. The separation of the Church of England from the church of Rome in the 16th century under Henry VIII did not have quite the repercussions in the scholastic field that were experienced by the continental Reformations. The secondary-school system in England had been strongly influenced by the Renaissance in the period preceding the reform, and about 300 grammar schools were already in existence. Nevertheless, the situation became precarious, for political reasons, under a succession of sovereigns.

Henry VIII included the schools in his policy of concentration and consolidation of power in the hands of the state. In 1548, under Henry's son Edward VI, the Chantries Act was passed, confiscating the estates of the church expressly for use in education; but the turmoil of the times, under the boy Edward and then his Roman Catholic sister Mary I, allowed the funds allocated to education to be diverted elsewhere. Many primary schools and grammar schools disappeared or retrenched their operations for lack of funds. Elizabeth I, however, succeeding to the throne in 1558, revived Henry VIII's educational policy; considerable sums were appropriated for education, even though it was not always possible to enforce the new provisions because of local opposition and some lack of concern on the part of the Anglican clergy.

The growth of a rich and prosperous mercantile class and the spread of Calvinist reforms through the Puritans in England and the Presbyterians in Scotland were also factors in the transformation of English education in the 16th and 17th centuries. Scholastic programs reflected changes in society: importance was given to English, to science, to modern languages (in particular French and Italian), and to sports, as is still the case in England today. The Puritan contribution was thus considerable, though often hindered by the traditional forces of the Anglican church and the old nobility.

The work of Sir Thomas Elyot and Roger Ascham

Sir Thomas Elyot, in *The Boke Named the Governour* (1531), wrote the first treatise in English that dealt specifically with education. He was interested in those who would have the future economic and political power in their hands. Though their education was to include the classics, it was to be supplemented by the needs of the new mercantile class—the national English language, manual arts, drawing, music, and all forms of sport. Elyot was obviously influenced by Erasmus.

Roger Ascham was close in thought to many of the English humanists. In *The Scholemaster* (1570) he underlined the importance of the English language (in spite of his being a professor of Greek) and proposed that it should be used in teaching the classical languages. He also believed that physical exercise and sport were important, not only for the nobility and the leisured classes but also for students and teachers. He was aware of the social changes in the country; and, observing with sadness the corruption of the new wealth, he was particularly chagrined to see students going to university not to gain culture but to prepare themselves for high offices of state.

Richard Mulcaster had 30 years of experience as an educator at St. Paul's School and at the Merchant Taylors School, a Latin secondary school maintained by the tailors' guild in London—and most famous of all the "guild schools." Mulcaster was in favour of efficient teacher training and of teachers being adequately paid. In agreement with some of the Lutheran educational reforms, he felt that schools should be open to all, including women, who should moreover have access to higher education. He is particularly remembered for his opposition to Italianate trends: "I love Rome, but London better. I favour Italy, but England more. I know the Latin, but worship the English."

Sir Francis Bacon was interested in education though it was not his main concern—his main concern being the championship of the scientific method and "sense" realism, or empiricism, in opposition to traditional Aristotelianism and Scholasticism. He was opposed to private tutors and felt that boys and youths were better off in schools and that their education should be geared to their social status and future activity. Schooling should aim at preparing statesmen and men of action as well as scholars and thus should include history, modern languages, and politics. Bacon himself had a passion for study not only for its utilitarian purposes but because of its being for him a true source of delight.

The French Reformation. Schools in 16th-century France were still largely under the control of the Roman Catholic church, as they had been in the Middle Ages. This traditional education faced opposition, however, both from Protestants and from reformers who had been influenced by the humanist principle of the primacy of the individual.

Reformers opposed to traditional education

François Rabelais was a great and original interpreter of humanistic ideals, and his views on education reflected this. He himself studied in various fields, from medicine to letters, and was passionately interested in all of them. His controversy with the Sorbonne, a remaining stronghold of medievalism and Scholasticism, was bitter; he satirized the school and the useless notions taught there in his novels *Pantagruel* (1532) and *Gargantua* (1534).

Rabelais's educational philosophy was entirely different from that of the medievalists—his being based on liberty of the pupil, in whom he had maximum faith. In *Gargantua* this cult of liberty was celebrated in the utopian Abbey of Thélème, where all could live according to their own pleasure but where the love of learning was so great that everyone was dedicated to it, getting much better results than those obtained at the medieval universities. And yet in the education of Gargantua and Pantagruel there were limits placed on liberty: Gargantua's day started at 4 in the morning; he studied all subjects, both literary and scientific; and this was alternated with play and pleasing diversions. The heavy program, however, was not a constriction because of Gargantua's delight in learning. The culture that Rabelais wanted for his two heroes was directly connected with the world in which they lived.

Gargantua and *Pantagruel* were perhaps among the first texts by a humanist in which not only the quadrivium but also scientific studies were enthusiastically proposed. There was nothing arid or abstract in Rabelais's approach

to nature, and in this context the classics also had a new flavour: ancient literature, no longer limited to Latin, Greek, and Hebrew but expanded to include Arabic and Chaldaic, could bring to light valuable knowledge that had been accumulated by the classical world.

Petrus Ramus, one of the most bitter critics of French medieval Aristotelianism, was an intelligent reformer of educational methods. His best-known treatises are *Aristotelicae animadversiones* (1543; "Animadversions on Aristotle") and *Dialecticae partitiones* (1543; "Divisions of Dialectic"), both condemned by royal decree; he also wrote two discourses on philosophy, *Oratio de studiis philosophiae et eloquentiae conjungendis* (1546; "Speech on Joining the Study of Philosophy with the Art of Speaking") and *Pro philosophica Parisiensis accademiae disciplina oratio* (1551; "Speech in Defense of the Philosophical Discipline of the Parisian Academy"), as well as *Ciceronianus,* published posthumously. In these works his criticism of traditional ways and of the degeneration of humanistic thought made him hated by all Roman Catholics, though not much better understood by Protestants; he died a Protestant victim of the massacre of St. Bartholomew. His program of study was fairly close to the traditional one, but his method was original, for he was concerned that the teacher should not suffocate the child with too many lessons and considered the child's autonomous activity important. He especially resented any pedagogy that relied on a blind appeal to authority; learning had to be utilitarian and issue from practice.

Michel de Montaigne (1533–92) was much influenced by his personal experience as a student. Though often critical of humanism, especially when it was misinterpreted and transformed into pedantic studies, he had great admiration for the classics and lacked the scientific interests of Rabelais or Ramus. Montaigne wrote specifically about education in two essays on the upbringing of children and on pedantry. Culture, he felt, had become imitation, often with no trace of originality left, whereas it should be a delight—not something a student is forced to assimilate but something to draw the student's participation. He was in favour of instruction by tutors capable of giving the student individual attention—the ideal tutor being one with a good mind rather than one filled with pedantic notions. He also believed in the importance of physical education and in a boy's being hardened to nature and to danger.

For Montaigne it was important not only to travel to foreign countries but also to stay there for a while, to learn languages and, even more, to learn about foreign customs and thus break out of the narrow limits of one's own province. There were many differences between Montaigne and Erasmus, but both were convinced that for the wise man there could be no geographic boundaries, for, through cultural diffusion, barriers would be broken down.

The Calvinist Reformation. The Protestant Reformer John Calvin was of French origin, but he settled in Geneva and made this Swiss city one of the most prominent centres of the Reformation. Unlike Luther, whose reforms were backed by princes hoping to gain greater political independence, Calvin was supported by the new mercantile class, which needed political and administrative changes for the purposes of its own expansion.

Calvin considered popular education important, but he was not an innovator. The theological academy he founded in Geneva in 1559 was modeled on Sturm's school in Strassburg, where Calvin had taught; it became distinguished under the directorship of Theodore Beza, an intelligent Reformer but unfortunately a very intolerant one, at least in theological matters. Calvin's influence on education was nevertheless felt in many of the European universities, even as far as England, where, in spite of Anglican opposition, the Puritans had gained a foothold.

Calvinist emphasis on universal literacy

Calvin was in favour of universal education under church control (the cost to be in large part borne by the community), but "universal" did not mean "democratic." Even if some form of instruction was to be given to everyone (so that everyone might in some measure read the Scriptures for himself, in good Calvinist tradition), very few individuals reached secondary or higher education, and of these only a minute percentage came from the working classes. Documents of the period show the steps taken to achieve the aim of universal education. In the Netherlands, the Calvinist Synod of The Hague in 1586 made provision for setting up schools in the cities, and the Synod of Dort in 1618 decreed that free public schools should be set up in all villages. In Scotland in 1560 John Knox, a disciple of Calvin and the leader of the Scottish Presbyterians, aimed at setting up schools in every community, but the nobility prevented this from actually being carried out. The major educational contributions of Calvinism were its diffusion to a larger number of people and the development of Protestant education at the university level. Not only was Geneva significant but also the universities of Leiden (1575), Amsterdam (1632), and Utrecht (1636) in the Netherlands and the University of Edinburgh (1582) in Scotland. The Puritan, or English Calvinist, movement was responsible for the founding of Emmanuel College at the University of Cambridge (1584).

The Roman Catholic Counter-Reformation. The religious upheaval, so important in northern Europe, also affected, though less violently, the Latin countries of southern Europe. If the new ferment in the Roman Catholic church was mainly directed at answering the Protestants, at times it also had something original to suggest. At the Council of Trent (1545–63) the Roman Catholic church tried to come to terms with the new political and economic realities in Europe.

Education was foremost in the minds of the leaders of the Counter-Reformation. The faithful were to be educated. For this, capable priests were needed, and, thus, seminaries multiplied to prepare the clergy for a more austere life in the service of the church. There was a flowering of utopian ideas, which should be remembered when trying to understand unofficial Catholic thought of the period. Writings such as *La città del sole* ("The City of the Sun"), by Tommaso Campanella, and *Repubblica immaginaria* ("The Imaginary Republic"), by Lodovico Agostini, are examples of this new vision of the church and of the duties of Christians. But if in the minds of the utopians this education was to be universal, it was in fact almost entirely directed at the ruling classes.

The growth of seminaries

The Society of Jesus, founded in 1534 by Ignatius Loyola, was not specifically a teaching order, but it was nevertheless very important in this field. The first Jesuit college was opened in Messina, Sicily, in 1548; by 1615 the Jesuits had 372 colleges, and by 1755, just 18 years before the suppression of the order, the number had risen to 728. (The society was not reestablished until 1814.) In *Ratio studiorum,* an elaborate plan of studies issued by the Jesuits in 1599, there is laid out an organization of these institutions down to the smallest details; an authoritarian uniformity was thus the rule in their colleges, and individual initiative was discouraged. The complete course of study took at least 13 years, divided into three periods: six or more years that included grammar and rhetoric, three years of philosophy, and four of theology. The teacher was thought of not only as an instructor but also as an educator and often a controller, for he was at the centre of a vast network of controls, in which those students considered promising also took part. Emulation was encouraged in the class, which was often divided into two groups to stimulate competition. These new techniques, as well as the Jesuits' efficient training of teachers, had good results, proof of this being the rapid increase in their colleges, which found greater favour than others started in the same period.

The legacy of the Reformation. The effects on education of a movement as complex and widespread as the Reformation were far-reaching. Perhaps its most original contribution was the extension of the idea of education at the elementary level. As a result, the vernacular language took on a new importance, and also the new pedagogy had to take account of the realities of the situation—namely, that the children brought into the new school network could not spend as much time on "useless" books, so that schoolwork had to be combined with learning a practical trade, which had not previously been considered a part of education. This, however, was to take several centuries to be implemented in practice. (E.Ge./J.Bo.)

European education in the 17th and 18th centuries

THE SOCIAL AND HISTORICAL SETTING

The Renaissance had been the beginning of a new era in history, which culminated in the 17th and 18th centuries in the development of the absolutist state everywhere but in England and Holland (and even in these states the issue was for some time in doubt). France, the Habsburg empire, England, and Russia became the leading powers in Europe. The absolutist state extended its control beyond the political and into the religious (with the creation of the established church) and into almost all other aspects of human life. Although the High and later Middle Ages had witnessed the growth of middle-class forces, the pattern of society still clearly bore the stamp of court life. The concentration of power determined this life, and the citizen and his possessions were more and more at the disposal of the aristocracy. The citizen was subject.

Influence of absolutism on education

Even in an absolutist state, however, education cannot be the sole privilege of the rich or the ruling classes, because an efficient absolutist state requires capable subjects, albeit bound to their social position. Elementary education for the middle classes thus developed in the 17th and 18th centuries, and more and more the state saw as its task the responsibility for establishing and maintaining schools. This tendency toward general education did not stem only from considerations of political expediency; it stemmed also from the desire to improve the world through education—making all areas of life orderly and subordinate to rational leadership. There was not only an inclination toward encyclopaedism and systemization of the sciences but also, in similar fashion, a tendency to set education aright by extensive school regulations.

In general, this distinction can be made between the 17th and the 18th centuries: in the 17th century the aim of education was conceived as a religious and rationalistic one, whereas in the 18th century the ideas of secularism and progress began to prevail. The 18th century is especially remembered for three leading reforms: teaching in the mother language grew in importance, rivaling Latin; the exact sciences were brought into the curriculum; and the correct methods of teaching became a pedagogic question.

The new scientism and rationalism. These social and pedagogic changes were bound up with new tendencies in philosophy. Sir Francis Bacon of England was one who criticized the teachers of his day, saying that they offered nothing but words and that their schools were narrow in thought. He believed that the use of inductive and empirical methods would bring the knowledge that would give man strength and make possible a reorganization of society. Therefore, he demanded that schools should be scientific workplaces in the service of life and that they should put the exact sciences before logic and rhetoric.

Another 17th-century critic of medievalism was René Descartes, but he did not proceed from empirical experience, as did Bacon; for him the only permanence and certainty lay in human reason or thinking (*cogito ergo sum,* "I think; therefore, I am"). The ability to think makes doubt and critical evaluation of the environment possible. A science based only on empiricism fails to achieve any vital, natural explanations but only mathematical, mechanistic ones of doubtful living use. Only what reason (*ratio*) recognizes can be called truth. Thus, education must be concerned with the development of critical rationality.

Knowledge conceived as thinking

Like Descartes, Benedict de Spinoza and Gottfried Wilhelm Leibniz also outlined rationalistic philosophic systems. Decisive for educational theory was their statement that knowledge and experience originate in thinking (not in sense impressions, which can provide only examples and individual facts) and that formal thinking categories should form the substance of education. They believed that the aim of education should be the mastery of thinking and judgment rather than the mere assimilation of facts.

The Protestant demand for universal elementary education. The schools that were actually developed fell short of these philosophically based demands. This is especially true of elementary education. In the Middle Ages, the grammar schools (especially for the education of the clergy) had developed, and the humanism of the Renaissance had strengthened this tendency; only those who knew Latin and Greek could be considered educated. For basic, popular education there were meagre arrangements. Although schools for basic writing and arithmetic had been established as early as the 13th and 14th centuries, they were almost exclusively in the towns; the rural population had to be content with religious instruction within the framework of the church. This changed as a result of Protestantism. John Wycliffe had demanded that everyone become a theologian, and Luther, by translating the Holy Scriptures, made the reading of original works possible. Everyone, he asserted, should have access to the source of belief, and all children should go to school. So it happened that church regulations of the 16th and 17th centuries began to contain items governing schools and the instruction of young people (mainly in reading and religion). At first, the Protestant schools were directed and supported almost entirely by the church. Not until the 18th century, following the general tendency toward secularization, did the state begin to assume responsibility for supporting the schools.

EDUCATION IN 17TH-CENTURY EUROPE

Central European theories and practices. It was while Europe was being shaken by religious wars and was disintegrating into countless small states that such writers as Campanella and Bacon dreamed their Utopias (*La Città del sole* and the *New Atlantis,* respectively), where peace and unity would be had through logical and realistic means. To even attempt realizing this dream, however, man needed suitable education. Both leading representatives of so-called pedagogic realism, Wolfgang Ratke and John Amos Comenius, were motivated by this ideal of world improvement through a comprehensive reform of the school system. Despite this common starting point, however, both were highly distinct personalities and, moreover, had divergent influences on the development of education and schools.

Realism in education

The pedagogy of Ratke. Ratke (1571–1635), a native of Holstein in Germany, journeyed to England, Holland, and through the whole of Germany and to Sweden expounding his ideas to the political authorities and finding considerable support. His plans for progressive reform failed for several reasons. First, political conditions during the Thirty Years' War were understandably not favourable for any kind of planning or reform of schools. Moreover, Ratke demonstrated little practical ability in executing his plans. Finally, Ratke's ideas were not free of exaggerations. He promised, for example, to be able to teach 10 languages in five years, each language in six months.

His ideas about the art of teaching are, nevertheless, of importance for the theory and practice of education. First, he believed that knowledge of things must precede words about things. This "sense realism" means that individual experience in contact with reality is the origin of knowledge; principles of knowledge follow, rather than precede, the study of specifics.

Second, everything must follow the order and course of what may be called human nature. In modern terms, one would say that a lesson should be designed with psychological conditions taken into consideration.

Third, he asserted that everything should be taught first in the mother language, the mother language being the natural and practical language for children and the one that allows them to concentrate wholly on the business at hand. Only when the mother language is fully commanded should a child attempt a foreign language; then special attention should be paid to speaking it and not merely reading it.

Fourth, Ratke emphasized what might now be called a kind of programmed learning. One piece of work should be fully completed before progress is made to the next piece, and there should be constant repetition and practice. The teacher's methods and the textbook program should agree and coincide.

Ratke's teacher-centred education

Fifth, there should be no compulsion. A teacher should not be a taskmaster. To strike a pupil was contrary to nature and did not help him learn. A pupil should be

brought to love his teacher, not hate him. On the other hand, all work was the teacher's responsibility. The pupil should listen and sit still. More generally, all children, without exception, should go to school, and no lessons should be canceled for any reason. There is, of course, a certain paradox in Ratke's views: there was to be no compulsion, and yet pupils were to remain disciplined and were not permitted to work independently.

As for curricula, Ratke suggested reading and writing in the native tongue, singing, basic mathematics, grammar, and, in the higher classes, Latin and Greek. The sciences had not yet appeared in his timetable. His demand that, above all, young people should be given instruction in the affairs of God is typical of the combination of rationalistic and religious education in the 17th century.

The pedagogy of Comenius. Comenius (1592–1670) was, even more than Ratke, a leading intellect of European educational theory in the 17th century. Born in Moravia, he was forced by the circumstances of the Thirty Years' War to wander constantly from place to place—Germany, Poland, England, Sweden, Hungary, Transylvania, and Holland—and was deprived of his wife, children, and property. He himself said, "My life was one long journey. I never had a homeland."

As a onetime bishop of the Bohemian Brethren, he sought to live according to their motto, "Away from the world towards Heaven." To prepare for the hereafter, Comenius taught, one should "live rightly"—that is, seek learned piety by living one's life according to correct principles of science and morality. Comenius' philosophy was both humanitarian and universalistic. In his *Pampaedia* ("Universal Education," discovered in 1935), he argued that "the whole of the human race may become educated, men of all ages, all conditions, both sexes and all nations." His aim was *pansophia* (universal wisdom), which meant that "all men should be educated to full humanity"—to rationality, morality, and happiness.

The curriculum of Comenius

Comenius realized that, to achieve *pansophia* by universal education, radical reforms in pedagogy and in the organization of schools were required, and he devised an all-embracing school system to meet this need. During infancy (up to six years of age), the child in the "mother school," or family grouping, would develop basic physical faculties. During the following period (seven to 12), the child would go to the "vernacular school," which was divided into six classes according to age and could be found in every town. The prime aim of these schools would be to develop the child's imagination and memory through such subjects as religion, ethics, diction, reading, writing, basic mathematics, music, domestic economy, civics, history, geography, and handicraft. This vernacular school formed the final stage of education for technical vocations. After this school would come the grammar school (or Latin school), which the pupils would attend during their youth (13–18) and which would exist in every town of every district. Through progressive courses in language and the exact sciences, the young people would be brought to a deeper understanding of things. Finally, the university (19–24) would be a continuation of this school. Every province ought to have one such university, whose central task would be the formation of willpower and powers of judgment and categorization. Over and above this four-tier school system Comenius also envisaged a "college of light," a kind of academy of the sciences for the centralized pooling of all learning. It is important to note, in this regard, that it was Comenius' stay in England (1641–42) that initiated discussions leading to the founding of the Royal Society (incorporated 1662). Furthermore, the German philosopher Leibniz, influenced by Comenius, founded the Berlin Academy, and similar societies sprouted elsewhere.

The Great Didactic (1657) sets forth Comenius' methodology—one for the arts, another for the sciences. Comenius believed that everything should be presented to the child's senses—and to as many senses as possible, using pictures, models, workshops, music, and other "objective" means. With proper presentation, the mind of the child could become a "psychological" counterpart of the world of nature. The mind can take in what is in nature if the method of teaching most akin to nature is used. For the upper age levels, he recommended that language study and other studies be integrated, and indeed he employed this scheme in his *Gate of Tongues Unlocked* (1631), a book of Latin and sciences arranged by subjects, which revolutionized Latin teaching and was translated into 16 languages. *The Visible World in Pictures* (1658), which remained popular in Europe for two centuries, attempted to dramatize Latin through pictures illustrating Latin sentences, accompanied by one or two vernacular translations.

The schools of Gotha. The zeal for reform on the part of such educators as Ratke and Comenius, on the one hand, and the interests of the ruling classes, on the other, led in the years after about 1650 to the publication of school regulations, free of church regulations. The circumstances in the central German principality of Gotha were typical. The duke, Ernst the Pious, commissioned the rector Andreas Reyher to compile a system of school regulations, which appeared in 1642 and is known historically as the *Gothaer Schulmethodus.* This was the first independent civil system of school regulations in Germany and was strongly influenced by Ratke. The most important points of these regulations were compulsory schooling from the age of five; division of the school into lower, middle, and higher classes; extension of the usual subjects (reading, writing, basic arithmetic, singing, and religion) to various other fields (natural history, local history, civics, and domestic economy); the introduction of textbooks (for reading and basic arithmetic), notably the first textbook of exact sciences for elementary schools, Reyher's own *Kurzer Unterricht von natürlichen Dingen* (1657; "Short Course on Natural Things"); and methodical instruction that, above all, emphasized the clarity of the lesson and the activity of the pupils.

Systematizing education

French theories and practices. In the second half of the 17th century Germany suffered from the aftereffects of the Thirty Years' War, whereas France under Louis XIV reached the zenith of political and military power. France's leadership was also demonstrated in the cultural field—including education. Some of the most important developments in France included the promotion of courtly education and the involvement of religious orders and congregations in the education of the poor.

Views of Montaigne

Courtly education. The rationalistic ideal of French courtly education can be seen foreshadowed in Montaigne's *Essays* (1580), in which the ideal man was described as having a natural, sensible way of life not deeply affected by the perplexities of the time but admitting of pleasure. He had a "correct" attitude toward the world and people, a certain spiritual freedom, and an independent judgment—all of which, in Montaigne's view, were more important than being steeped in knowledge. "As lamps are extinguished from too much oil, so is the mind from too much studying." Montaigne came from a merchant family that aspired to nobility, and thus there is a certain fashionable elitism in his views; he held, among other things, that courtly education succeeds best when the pupil studies under a private tutor.

This ideal, rather unlike the ideal of the learned and humanistic Renaissance man, became important in 17th-century France, especially after mid-century and the rise of the court of Louis XIV. The education of the would-be versatile and worldly-wise gentleman was furthered not only by the continuation of the institution of private tutoring but also by the establishment of schools and academies for chevaliers and nobles, in which the emphasis was on such subjects as deportment, modern languages, fencing, and riding. It was most emphatically an example of class education, designed for the nobility and higher military and not for any commoners.

The teaching congregations. In the countries, such as France, that remained Catholic, the Roman church retained control of education, and indeed, as monarchy became more absolute, so largely did the authority of the church in matters of education. In France, practically all schools and universities were controlled by so-called teaching congregations or societies, the most famous and powerful of which during the first half of the 17th century was the Society of Jesus. By mid-century the Jesuits

had 14,000 pupils under instruction in Paris alone; and their colleges (not including universities) all over the land numbered 612.

Jesuit education

It was their successful teaching and comparatively mild discipline that caused the Jesuit schools to attract thousands of pupils. "They are so good," said Bacon of the Jesuit teachers in his *Advancement of Learning,* "that I wish they were on our side." The curriculum was purely classical, but importance was attached to spacious, well-adapted buildings and amenities designed to make school life interesting. In general, however, the religious and international conflicts did great harm to education, which suffered much because those kings and religious factions that gained power in France (as elsewhere) used the schools to propagate their cause, discarding teachers not of the approved persuasion. Moreover, the schools continued largely to ignore the new directions of men's minds; in the universities staffed by Jesuit fathers, medieval Scholasticism, though purged of the formalistic excesses that had degraded it, was fully restored. Schools and universities declined for the most part to contemplate any enlargement of the frontiers of knowledge and were too often deeply involved in the religious conflicts of the time. The University of Paris in particular remained distracted throughout the 17th century by theological dissensions—in at least one instance as a result of the rivalry that ensued after the Jesuits had effected a footing at Clermont College.

Aside from the Jesuits, the most important teaching congregations in France were the Bérullian Oratory, or Oratorians, and the Jansenists of Port-Royal. The former, founded in 1611 and soon to open a number of schools and seminaries for young nobles, was composed of priests—but priests more liberal and rationalist than was common for the times. They offered instruction not only in the humanities but also in history, mathematics, the natural sciences, and such genteel accomplishments as dancing and music and, though continuing to use Latin in instruction, promoted also the use of the vernacular French in the initial years of their curriculum. They tended indeed to be drawn to the ideas of Descartes, to a faith based on reason. When in 1764 the Jesuits were banned from France, their teaching positions were largely assumed by Oratorians.

The Jansenists

More famous than the schools of the Oratorians, though enjoying a briefer career, were the Little Schools of Port-Royal. Their founder was Jean Duvergier de Haurame, better known as the abbot of Saint-Cyran, who was one of France's chief advocates of Jansenism, a movement opposed to Jesuitry and Scholasticism and favouring bold reforms of the church and a turn to a certain Pietism. About 1635 Saint-Cyran, with the help of some wealthy, influential Parisians, succeeded in gaining control of the convent of Port-Royal, near Versailles. There the Jansenist group began about 1637 to educate a few boys, and by 1646 it had established the Little Schools of Port-Royal in Paris itself. Their curriculum was similar to that of the Oratorians, though excluding dancing, and was celebrated for its excellence in French language and logic and in foreign languages. Influenced by Descartes's rationalistic philosophy, the Jansenists theorized that learning has a "natural" order and should begin with what is familiar to the child: thus, a phonetic system of teaching reading was used; all instruction was in French, not Latin; student compositions were directed toward topics drawing on one's own experiences or toward subjects in one's current reading. Involved in political struggles with the Jesuits, who were still influential at court, the Jansenists were fated to have all their schools closed down by 1660, but their theories and practices were widely adopted and became extremely influential.

Female education. During the century, the education of girls was not entirely neglected, and France was notable for its efforts. Mme de Maintenon, for instance, had been a pupil of the Ursuline nuns in Paris and then a governess at the court of Louis XIV before she was wedded to the king in 1684. From her royal vantage point, she took upon herself the founding of a school in 1686 at Saint-Cyr near Versailles—a higher school principally for orphan girls descended from noble families. Besides such basic subjects as reading and writing, the girls were prepared for their future lives as wives and mothers or as members of genteel professions. In 1692 this school was taken over by the Augustinian nuns. Another important worker in the field of female education was St. Jane Frances de Chantal, who, together with her father confessor, St. Francis de Sales, founded in 1610 the order of the Visitandines, a group dedicated to charitable work and the religious education of women.

Views of Fénelon

François de Salignac de La Mothe-Fénelon, archbishop of Cambrai and noted theologian and writer, is especially known for his views on the education of girls. In his *Traité de l'éducation des filles* (1687; "Treatise on the Education of Girls") he remarked on the importance of women in improving the morals of society and went on to express his thoughts about girls' education. Because girls, he believed, are meant to fulfill roles as housewives and mothers, they should pursue religious and moral education rather than scholarly learning. They should learn reading and writing, basic mathematics, history, music, needlework, Latin (because it is the church language), but no modern languages, since they tend to moral corruption. Education, he maintained, should make the lady of the house both Christian and accomplished, neither ignorant nor précieuse.

English theories and practices. The 17th century in England (up to the Glorious Revolution of 1688–89) was one of argument over religious and political settlements bequeathed by Queen Elizabeth I; the period was one characterized by the confrontation of two different world-views—on one side the royalist Cavaliers and on the other side the Puritans. The division was reflected in education.

The Puritan Reformers. In the Anglo-American world the Reformation came about in the form of Calvinism—"Puritan" being the derisory name for strict Calvinists. Their ideals were sober, practical behaviour, careful management, thrift, asceticism, and the rejection of hedonistic pleasures of life. Many of the educationists who sought this Puritan ideal were followers of the reform plans of Comenius. Samuel Hartlib, a Polish merchant residing in England who was friend, publisher, and patron of Comenius, tried to interest Parliament in the idea of popular education; his treatise *London's Charity Enlarged* (1650) proposed that a grant be made for the education of poor children, all in the interest of general social betterment. The Committee for Advancement of Learning, which he founded in 1653, was the impulse and model for later educational associations. In general, his ideas for reform included the introduction of agricultural schools and the state organization of the educational system, as well as the establishment of general elementary education.

The name of John Dury stands close to those of Comenius and Hartlib. In 1651 appeared his book *The Reformed School,* in which he proposed teaching societies in England much like the teaching congregations in France. Indeed, he was particularly insistent that control of education be in the hands not of a regimentizing state but of free educational organizations. He was also concerned about teaching youth the useful arts and sciences so that they might "become profitable instruments of the Commonwealth." From him, too, stemmed the draft of a nursery school; thus, he can be regarded as the first representative of infant teaching in England.

The humanism of John Milton

The most renowned of the Puritan intellectuals, John Milton, was more concerned with the education of "our nobler and our gentler youth" than with the education of common boys. *Of Education* (1644), written at the request of Hartlib, was one of the last in the long line of European expositions of Renaissance humanism. Milton's aim was the traditional aim, the molding of boys into enlightened, cultivated, responsible citizens and leaders. His proposed academy, which would take the place of both secondary school and college, was to concentrate on instruction in the ancient classics, with due subordination to the Bible and Christian teaching. Milton also emphasized the sciences, and physical and martial exercise had a place in his curriculum as well.

Royalist education. Frequently opposed to Puritanism on educational as well as political grounds were the royalists and supporters of the nobility. In education, their

views went back to Elyot and Ascham in the 16th century, who had written so persuasively about the education of gentlemen in the tradition of the so-called courtesy books. Influenced by these few English forerunners and also by Montaigne were James Cleland (*The Institution of a Young Nobleman,* 1607) and Henry Peacham (*The Compleat Gentlemen,* 1622). In the view of the latter, an extreme royalist, "Fashioning him [the pupil] absolute in the most necessary and commendable Qualities concerning Minde and Body to country's glory" was the overriding aim of education; the table of contents of *The Compleat Gentlemen* exhibits the variety of interests of an ideal gentleman or noble—cosmography, geometry, poetry, music, sculpture, drawing, painting, heraldry, and so on. John Gailhard (*The Compleat Gentleman,* 1678), another writer in the same tradition, can be said to have anticipated John Locke's empiricism (see below) when he wrote that "the nature of Youth is like Wax by fire, or a smooth table upon which anything can be written."

The academies. The beginning of academies for the promotion of philosophy, arts, or sciences can be traced to the early Renaissance, particularly in Italy and France. The Platonic Academy in Florence, cited earlier in this article, was one of the most noted of speculative societies. The first scientific academies belong to the 16th century: in 1560, for instance, the Academia Secretorum Naturae ("Secret Academy of Nature") was founded in Naples; in 1575 Philip II of Spain founded in Madrid the Academy of Mathematical Sciences. Then, in 1617, the first German academy, Fruchtbringende Gesellschaft ("Productive Society"), was founded at Weimar with the expressed purposes of the purification of the language and the cultivation of literature. A number of other academies were founded throughout Europe.

The Royal Society and the Academy of Sciences

It was in the 17th century that the two preeminent scientific academies were founded. Both the English Royal Society and the French Academy of Sciences began as informal gatherings of famous men. The "invisible college" of London and Oxford had its first meetings in 1645; it was incorporated as the Royal Society in 1662. In Paris, a group of men including the philosophers Descartes and Pascal started private meetings almost at the same time. In 1666 they were invited by the economic minister Jean-Baptiste Colbert to meet in the royal library. In 1699 the society was transferred to the Louvre under the name of the Academy of Sciences. The French Academy also started as a private society of men of letters some five years before its incorporation in 1635 under the patronage of Cardinal de Richelieu. In the 18th century, the fame and achievements of these English and French academies became internationally recognized, and many other European countries started to found their own national academies.

EDUCATION IN 18TH-CENTURY EUROPE

In the 18th century the theories and systems of education were influenced by various philosophical and social trends. Among these were realism, which had its origins in Ratke and Comenius, among others, and also Pietism, which derived principally from Philipp Jakob Spener and August Hermann Francke in the late 17th and early 18th centuries (see below). Another trend was the far-reaching rationalistic and humanitarian movement of the Enlightenment, best seen in the pedagogical views of Locke, in the upsurge of philanthropy, and in Denis Diderot's *Encyclopédie,* a comprehensive system of human knowledge in 28 volumes (1751–72). Also important was naturalism, of which Jean-Jacques Rousseau can be regarded as the main representative.

Locke's theory of knowledge

Education during the Enlightenment. *John Locke's empiricism and education as conduct.* The writings of the late 17th-century empiricist John Locke on philosophy, government, and education were especially influential during the Enlightenment. For education, Locke is significant both for his general theory of knowledge and for his ideas on the education of youth. Locke's empiricism, expressed in his notion that ideas originate in experience, was used to attack the doctrine that principles of reason are innate in the human mind. In *An Essay Concerning Human Understanding* (1690) Locke argued that ideas come from two "fountains" of experience: sensation, through which the senses convey perceptions into the mind, and reflection, whereby the mind works with the perceptions, forming ideas. Locke through of the mind as a "blank tablet" prior to experience, but he did not claim that all minds are equal. He insisted, in *Some Thoughts Concerning Education* (1693), that some minds have a greater intellectual potential than others.

For education, Locke's empiricism meant that learning comes about only through experience. Education, which Locke felt should address both character and intellect, is therefore best achieved by providing the pupil with examples of proper thought and behaviour, by training the child to witness and share in the habits of virtue that are part of the conventional wisdom of the rational and practical man. Virtue should be cultivated through proper upbringing, preparatory to "studies" in the strict sense. The child first learns to do through activity and, later, comes to understand what has been done. The intimacy between conduct and thinking is best illustrated in the title of Locke's *Of the Conduct of the Understanding,* written as an appendix to his *Essay.* There it is clear that understanding comes only with careful cultivation and practice; this means that understanding not only involves conduct but is itself a kind of conduct. If the child and the tutor share a kind of conduct, then the child will have learned the habits of character and mind that are necessary for education to continue.

Vico's emphasis on imagination

Giambattista Vico, critic of Cartesianism. Like Locke, the Italian philosopher Giambattista Vico believed that human beings are not innately rational; he argued, however, that understanding results not through sense perception but through imaginative reconstruction. Although Vico's ideas were not widely known in the 18th century, the importance of his work for the history of philosophy and education has been increasingly recognized since the late 1960s. Vico was professor of rhetoric at the University of Naples from 1699 to 1741. His best-known work is *New Science* (1725), in which he advanced the idea that human beings in their origins are not rational, like philosophers, but imaginative, like poets. The relation between imagination and reason in *New Science* is suggestive for educational theory: civilized human beings are rational, yet they came to be that way without knowing what they were doing; the first humans created institutions literally without reason, as poets do who follow their imagination rather than their reason. Only later, after they have become rational, can human beings understand what they are and what they have made. Vico's idea that early humans were nonrational and childlike prefigured Rousseau's primitivism and his conception of human development (see below); and the importance Vico accorded to imagination foreshadowed the place that feeling was to have in 19th-century Romantic thought.

De Nostri Temporis Studiorum Ratione (1709; "On the Study Methods of Our Time") defended the humanistic program of studies against what Vico took to be an encroachment by the rationalistic system of Descartes on the educational methods proper for youth. Vico asserted that the influential Cartesian treatise *The Port-Royal Logic,* by the Jansenists Antoine Arnauld and Pierre Nicole, inverted the natural course by which children learn by insisting on a training in logic at the beginning of the educational process. He argued, instead, that young people need to have their mental powers developed and nourished by promoting their memories through the study of languages and enhancing their imaginations through reading poets, historians, and orators. Young minds first need the kind of reasoning that common sense provides. Common sense, acquired through the experience of poets, orators, and people of prudence, teaches the young the importance of working with probabilities prior to an education in logic. To train youth first in logic in the absence of common sense is to teach them to make judgments before they have the knowledge necessary to do so. Vico's aim was to emphasize the importance of practical judgment in education, an echo of the ideals of Locke and a prefiguring of Rousseau and the 19th-century reformer Johann Heinrich

Pestalozzi. Outside of Italy, among those who were most influenced by *New Science* were Joseph de Maistre in the late 18th century and Victor Cousin and Jules Michelet in the 19th century.

The condition of the schools and universities. The school system became more and more in the 18th century an ordered concern of the state. Exponents of enlightened absolutism, as well as parliamentarians, recognized that the subject was of more use to the state if he had a school education. Ideally, there was to be compulsory schooling everywhere, but of course in practice the ideal was scarcely reached anywhere. The state also recognized that worthwhile school instruction depended on the standard of education of teachers: thus, the first teachers' colleges were established. But admittedly the standard of education of teachers was fairly poor. The teaching profession still did not provide a living wage, for which reason can be read from a regulation of 1736:

> If the teacher is a workman he can already support himself; if he is not, then he is hereby allowed to go to work for daily wages for 6 weeks at harvest-time (*Principia regulativa,* clause 10).

Ever since the 16th century the universities had suffered a decline, mainly as a result of religious wars. Progress in the exact sciences was accomplished under government support in the academies of science, not in the universities, which became more and more training institutions for higher civil servants. There was, however, a notable change for the better, at least in Germany.

Halle, the first modern university

The year 1694 saw the foundation of the University of Halle, which has been described as the first real modern university. It originated in a *Ritterschule,* or "knight's school," imitative of the schools for chevaliers in France, and in 1694 the Holy Roman emperor Leopold I granted it a charter. The primary object in founding a university in Halle was to create a centre for the Lutheran party; but its character, under the influence of its two most notable teachers, the philosophers Christian Thomasius and Francke, soon expanded beyond the limits of this conception. Thomasius was the first to set the example—soon after followed by all the universities of Germany—of lecturing in the vernacular instead of the customary Latin; this was a declaration of war against Scholasticism. Francke, as the founder of the Pietistic school, exercised great influence. Throughout the whole of the 18th century Halle was the leader of academic thought and advanced theology in Protestant Germany, although sharing that leadership, after the middle of the century, with the University of Göttingen (founded 1737). With Göttingen, another important contribution was made by the revival of classical studies and the creation of a faculty of philosophy distinct from that of theology. This was designed not only to advance scholarship but also to train teachers. Halle itself established the first chair of educational theory.

The background and influence of Pietism. The dispute over the correct religious dogma, fought for almost 200 years with the utmost strength, controversy, and academic subtlety and reaching its terrible culmination in the Thirty Years' War, led to a certain ill feeling against dogmatically sanctioned religious revelation. There was a widespread trend toward secularization. Everywhere, there was a clear tendency to free belief from dogmatic quarrels. The search for a new belief took generally two different paths. One

Rationalistic versus devotional experience

wanted to base belief in man's reason; the other wanted a godliness of the heart. For one line of thought, belief was a postulate of omnipotent human reason; for the other, man, corrupted by original sin, was to be saved only by simple belief in God's grace. The one path turned to the religious understanding of the Enlightenment; the other followed the subjective, mystical, zealous devoutness of Pietism. Such a movement away from the institutionalized church, away from the established church, and toward an intensified faith was evident in France within Roman Catholicism in the form of Jansenism and Quietism. In England it was clearly evident in certain forms of Puritanism and in Independent movements and Quakerism. In Germany it was evident in Pietism.

Pietism was a Protestant movement of renewed faith that became popular from about 1675 to 1740, though it remained residually influential even into the 19th century. Its spiritual centres were in Württemberg, among the Moravian Brethren, and above all in Halle. Pietism was principally opposed to dogmatic Protestant orthodoxy, which usually included impatience and polemics against other beliefs. Pietism, on the contrary, stood for the renewal of importance of the individual prayer and for humility. The experiences of belief were to be based less in the acceptance of fixed conditions of belief and more in a mystical, personal submersion in feelings. According to standard Protestant theory, salvation could be hoped for only by the suppression of the corrupted individuality and by waiting for the grace of God to show one the way. From this came the Pietists' inclination to turn away from the world with its temptations (*e.g.,* the theatre, dancing, games, and other enjoyments). The uneasiness that they felt toward church institutionalization led to their splitting into numerous separatist groups; their subjective certainty about their belief led to a certain arrogance; and finally their seclusion led often to a joyless and moralizing way of life.

Although the founder of German Pietism is considered to be Spener, who established several private devotional gatherings (*collegia pietatis*) for Bible study in Frankfurt am Main and elsewhere, he was important for education only in the sense that he fashioned a spirit or concept in which education could be conducted—a concept that would subordinate all education to a simple Christian faith. This concept was realized mainly by his follower Francke.

August Hermann Francke. Francke, after service as a grammar-school teacher and priest in Leipzig, Lübeck, Hamburg, and Erfurt, was, through Spener's recommendation, given a post at the University of Halle in 1691, at the same time assuming the post of parish priest nearby. Motivated by the sad conditions of neglect in his parish, he quickly devoted himself to practical pastoral duties. In 1695 he instituted a vernacular school for the poor, popularly called the "ragged school," whose purpose was

The schools of Halle

that the children should be led to a living knowledge of God and Christ and to a rightly accomplished Christianity. Through his activity and eloquence Francke won several charitable patrons for his school, and the institution quickly expanded. After the school for the poor came the establishment of an elementary school for children of fee-paying burghers, then an orphanage, and lastly a *Pädagogium,* or boarding school, for the sons of nobility. Because Francke felt a lack of suitable teachers for his schools, he subsequently established two teachers' seminaries, *seminarium praeceptorum* and *seminarium selectum* (for teachers in higher schools). In 1697 there followed a Latin grammar school and in 1698, even if short-lived, a *gynaeceum,* a school for the daughters of nobility. To the whole complex of Halle's institutions (known collectively as the Halle Foundation) there also belonged a bookshop with a publishing house and press, a very profitable chemistry laboratory, as well as four agricultural properties, a Bible institution, and an office for sending evangelical missions abroad. These institutions flourished, and about 1750 they were more and more brought under the control of the state.

Francke's main concern was ministerial work in the spirit of Pietism and not systematic educational theorizing. His educational aims were religious and at the same time practical. He himself paraphrased it as "true godliness and Christian wisdom"—true godliness meaning a pious, moral, devout life, and Christian wisdom referring to an

Strictness and realism in Pietistic schools

ability to work hard according to the Protestant ethic. Francke's style of education went along with this aim: the corrupted willfulness of man must be broken, not through severe punishment but through "loving reproaches," a close supervision of the pupils, and a schooled and regimented care of the spirit. Games and childlike exuberance have no place in the system; thus, education had a joyless and moralizing effect.

The harsh demands and regimentation are shown, for instance, in the daily timetable and the syllabus. The children arose at 5:00 AM; there was almost continuous instruction with frequent Bible reading and religious

lessons until 7:00 in the evening. The grammar school had lessons in reading, writing, basic mathematics, catechism, the Holy Scriptures, Latin, Greek, Hebrew, optionally another Oriental language, geography, history, mathematics (including astronomy and geometry), botany, zoology, mineralogy, anatomy, and theology, as well as lathework, glass polishing, field trips to observe trades, factory work, horticulture, and so on. These latter subjects were counted as "recreation." The pansophic idea of Comenius was being followed here, in the sense that there was to be an all-encompassing education. It is worth noting that Francke was actually trying to inject realism into education—promoting, as he did, scientific subjects, lessons in manual skills, planned field trips, and even the reading of newspapers in the classroom.

Johann Julius Hecker. Julius Hecker came to Halle shortly before Francke's death in 1727 and became a teacher in the *Pädagogium.* In 1739 he was summoned by Frederick I of Prussia to Berlin, where he established a six-year *Realschule,* or "realist school," designed to prepare youth for the Pietistic and Calvinistic ideal of hard work and, especially, for the new technical and industrial age that was already dawning in countries such as England and France. Godliness was to be combined with a realistic and practical way of life. As early as 1699 Francke had conceived the idea of a school for children who were not meant for scholarship but who could serve usefully in commercial pursuits or administration; and in 1739 one of his teachers, Christoph Semler, published a pamphlet proposing such a "mathematical and mechanical *Realschule.*" It was Hecker's fortune to put these plans into realization. His school included, among other things, classes for architecture, building, manufacturing, commerce, and trade. Both the exact sciences and manual skills were in the curriculum. A room for natural-history specimens, geographic maps, and realia was set aside for the illustration of lessons. Schools like Hecker's were gradually opened in other cities. In the 19th century courses were extended to nine years, and such an institution was renamed *Oberrealschule,* or "higher realist school"; henceforth it was one of the main types of German secondary education. Hecker also compiled the general school regulations (1763) that formed the main outlines of the Prussian school system.

Realschule, or "realist school"

The background and influence of naturalism. Pietists emphasized Christian devotion and diligence as paths to the good life; Enlightenment thinkers focused on reason and clear thinking as the sensible way to happiness. Rousseau and his followers were intrigued by a third and more elusive ideal: naturalism. Rousseau, in his *A Discourse on Inequality,* an account of the historical development of the human race, distinguished between "natural man" (man as formed by nature) and "social man" (man as shaped by society). He argued that good education should develop the nature of man. Yet Rousseau found that mankind has not one nature but several: man originally lived in a "pure state of nature" but was altered by changes beyond control and took on a different nature; this nature, in turn, was changed as man became social. The creation of the arts and sciences caused man to become "less pure," more artificial, and egoistic, and man's egoistic nature prevents him from regaining the simplicity of original human nature. Rousseau is pessimistic, almost fatalistic, about changing the nature of modern man.

Émile, his major work on education, describes an attempt to educate a simple and pure natural child for life in a world from which social man is estranged. Émile is removed from man's society to a little society inhabited only by the child and his tutor. Social elements enter the little society through the tutor's knowledge when the tutor thinks Émile can learn something from them. Rousseau's aim throughout is to show how a natural education, unlike the artificial and formal education of society, enables Émile to become social, moral, and rational while remaining true to his original nature. Because Émile is educated to be a man, not a priest, a soldier, or an attorney, he will be able to do what is needed in any situation.

Rousseau's educational ideas in *Émile*

The first book of *Émile* describes the period from birth to learning to speak. The most important thing for the healthy and natural development of the child at this age is that he learn to use his physical powers, especially the sense organs. The teacher must pay special attention to distinguishing between the real needs of the child and his whims and fancies. The second book covers the time from the child's learning to speak to the age of 12. Games and other forms of amusement should be allowed at this age, and the child should by no means be overtaxed by scholarly instruction at too early an age. The child Émile is to learn through experience, not through words; he is to bow not to the commands of man but to necessities. The third book is devoted to the ages from 12 to 15. This is the time of learning, not from books of course but from the "book of the world." Émile must gain knowledge in concrete situations provided by his tutor. He learns a trade, among other things. He studies science, not by receiving instruction in its facts but by making the instruments necessary to solve scientific problems of a practical sort. Not until the age of 15, described in the fourth book, does Émile study the history of man and social experience and thus encounter the world of morals and conscience. During this stage Émile is on the threshold of social maturity and the "age of reason." Finally, he marries and, his education over, tells his tutor that the only chains he knows are those of necessity and that he will thus be free anywhere on earth.

The final book describes the education of Sophie, the girl who marries Émile. In Rousseau's view, the education of girls was to be similar with regard to naturalness, but it differed because of sexual differences. A girl cannot be educated to be a man. According to Rousseau, a woman should be the centre of the family, a housewife, and a mother. She should strive to please her husband, concern herself more than he with having a good reputation, and be satisfied with a simple religion of the emotions. Because her intellectual education is not of the essence, "her studies must all be on the practical side."

At the close of *Émile,* Rousseau cannot assure the reader that Émile and Sophie will be happy when they live apart from the tutor; the outcome of his experiment is in doubt, even in his own mind. Even so, probably no other writer in modern times has inspired as many generations as did Rousseau. His dramatic portrayal of the estrangement of natural man from society jolted and influenced such contemporary thinkers as Immanuel Kant and continues to intrigue philosophers and social scientists. His idea that teachers must see things as children do inspired Pestalozzi and has endured as a much-imitated ideal. Finally, his emphasis on understanding the child's nature had a profound influence by creating interest in the study of child development, inspiring the work of such psychologists as G. Stanley Hall and Jean Piaget.

The Sensationists. A group of French writers contemporary with Rousseau and paralleling in some ways the thought of both Rousseau and Locke are known as the Sensationists, or, sometimes, the Sensationist psychologists. One of them was Étienne Bonnot de Condillac, who, along with Voltaire, may be said to have introduced Locke's philosophy to France and established it there.

In the *Treatise on Sensations* (1754) Condillac imagined a statue organized inwardly like a man but animated by a soul that had never received an idea or a sense impression. He then unlocked its senses one by one. The statue's power of attention came into existence through its consciousness of sensory experience; next, it developed memory, the lingering of sensory experience; with memory, it was able to compare experiences, and so judgment arose. Each development made the statue more human and dramatized Condillac's idea that man is nothing but what he acquires, beginning with sensory experience. Condillac rejected the notion of innate ideas, arguing instead that all faculties are acquired. The educational significance of this idea is found in Condillac's *An Essay on the Origin of Human Knowledge* (1746), where he writes of a "method of analysis," by which the mind observes "in a successive order the qualities of an object, so as to give them in the mind the simultaneous order in which they exist." The idea that there is a natural order which the mind can learn to follow demonstrates Condillac's naturalism along with

Views of Condillac

his sensationism. Condillac does not begin his work *Logic* (1780) with axioms or principles; rather, he writes, "we shall begin by observing the lessons which nature gives us." He explains that the method of analysis is akin to the way that children learn when they acquire knowledge without the help of adults. Nature will tell man how to know, if he will but listen as children "naturally" do. Thus the way in which ideas and faculties originate is the way of logic, and to communicate a truth is to follow the order in which ideas come from the senses.

Claude-Adrien Helvétius, a countryman of Condillac's who professed much the same philosophy, was perhaps even more insistent that all human beings lack any intellectual endowment at birth and that despite differing physical constitutions each person has the potential for identical passions and ideas. What makes people different in later life are differing experiences. Hypothetically, two men brought up with the same chance experiences and education would be exactly the same. From this it followed, in education, that the teacher must attempt to control the environment of the child and guide his instruction step by step. Helvétius was, perhaps, unique in joining such a strong belief in intellectual equalitarianism with the possibility of a controlling environmentalism.

The Rousseauists. Rousseau left behind no disciples in the sense of a definite academic community, but hardly a single theorist of the late 18th century or afterward could avoid the influence of his ideas. One of those influenced was the German Johann Bernhard Basedow, who agreed with Rousseau's enthusiasm for nature, with his emphasis on manual and practical skills, and with his demand for practical experience rather than empty verbalism. The teacher, in Basedow's view, should take pains over the clearness of the lesson and make use of the enjoyment of games: "It is possible to arrange nearly all playing of children in an instructive way." In another respect, however, the contrast between Rousseau and Basedow could not be sharper; Basedow tended to force premature learning and overload a child's capabilities. A foreign language, for instance, was to be learned in six months. He promoted, in general, a pedagogic hothouse atmosphere. Basedow was perhaps influenced by his seven-year-old daughter, who was put forward as a wonder child with extraordinary knowledge. He established an experimental school called a *Philanthropinum,* in Dessau, which lasted from 1774 to 1793.

Kant on pedagogy

Kant referred to Rousseau's influence on him. He dealt specifically with pedagogy only within a lecture he gave as holder of the chair of philosophy in Königsberg; the main features of the lecture were collected in a short work, *Über Pädagogik* (1803; "On Pedagogy"). In it he asserted, "A man can only become a man through education. He is nothing more than what education makes him." Education should discipline man and make him cultured and moral; its aim is ultimately the creation of a happier mankind. In general, Kant agreed with Rousseau's education according to nature; but, from his ethical posture, he insisted that restraints be put on the child's passionate impulses and that the child even be taught specific maxims of conduct. The child must learn to rule himself and come to terms with the twin necessities of liberty and constraint, the product of which is true freedom.

> Children should be educated, not with reference to the present conditions of things, but rather with regard to a possibly improved state of the human race—that is, according to the ideal of humanity and its entire destiny. (From *Über Pädagogik.*)

The influence of nationalism. The Enlightenment was cosmopolitan in its effort to spread the light of reason, but from the very beginning of the age there were nationalistic tendencies to be seen in varying shades. Although Rousseau himself was generally concerned with universal man in such works as *The Social Contract* and *Émile,* his *The Government of Poland* (1782) did lay out a proposal for an education with a national basis, and generally his ideas influenced the nationalistic generation of the French Revolution of 1789.

National views of the Philosophes

France. The real starting point generally of national pedagogic movements was in France. It perhaps began with the Philosophes, the rationalists and liberals such as Voltaire and Diderot, who emphasized the development of the individual through state education, not as a means, of course, of adjusting to the state and its current government but as a means of creating critical, detached, responsible citizens. The Marquis de Condorcet was closely connected with this line of thought. For him man was by nature good and capable of never-ending perfection, and the goal of education should be the "general, gradually increasing perfection of man." He drafted a democratic and liberal but at the same time somewhat socialist concept of school policy: there should be a uniform structure of public education and equal chances for all; ability and attainment should be the only standards for selection and careers; and private interests should be prevented from having influence in the educational system. An educational concept so rationalistic in its aims and with such a democratic and liberal structure cannot be narrowly nationalistic; it is cosmopolitan. But Condorcet was nationalistic insofar as he wanted "to show the world at last a nation in which freedom and equality for all was an actuality." He was, in fact, a strong supporter of the Revolution.

Many of the Rousseauists were nationalistic in a somewhat different way. They believed in a kind of "moral patriotism." They distrusted state-controlled nationalism and favoured instead a virtuous, patriotic citizen who experienced spontaneous feelings for his nation. Proper development in the family setting and in school would lead to the mastery of everyday situations and would naturally lay the foundations for this true nationalism.

Some of the French revolutionists, particularly Jacobins such as Robespierre and Saint-Just who were associated with the period of the Terror (1793–94), were concerned with an education for the revolutionary state, an education marked by an enmity toward the idea of scholarship for its own sake and by state control, collectivism, the stressing of absolute equality, and the complete integration of all. What is good is decided by the collective "people." Thus, it could be said that the Jacobins favoured a complete politicalization of educational practice and theory.

Benevolent despotism and education

National education under enlightened rulers. The absolutism of the 18th century has often been called "benevolent despotism," referring to the rule of such monarchs as Frederick II the Great of Prussia, Peter I the Great and Catherine II the Great of Russia, Maria Theresa and Joseph II of Austria, and lesser figures who were presumably sufficiently touched by the ideas of the Enlightenment to pursue social reforms. Their reforms were limited, however, and usually did not include anything likely to upset their sovereignty. Thus, they were often willing to improve education for middle-class persons useful in civil service and other areas of state administration, but they were often chary of educating the poor. That risked upsetting the social order.

Frederick the Great, however, issued general school regulations (1763) establishing compulsory schooling for boys and girls from five to 13 or 14 years of age. His minister Freiherr von Zedlitz founded a chair of pedagogy at Halle (1779) and generally planned for the improved education of teachers; he supported the founding of new schools and the centralization of school administration under an *Oberschulkollegium,* or national board of education (1787); and one of his colleagues, Friedrich Gedike, was instrumental in introducing the school-leaving examination for university entrance, the *Abitur,* which still exists.

The guarded though increasingly liberal attempt by benevolent despots to nationalize and expand education is well illustrated by the events in Russia. Until the 18th century, schools in Russia were founded by ecclesiastical organizations (monasteries), the clergy (priests, deacons, readers), and private persons (boyars, or lower-level aristocrats). Boys were taught reading, writing, arithmetic, singing, and religion. A system of state-owned schools was started by Peter the Great as a state organization for purposes of administration and for the development of mining and industry. Peter did not intend to promote the Orthodox faith or formal classical learning, whether Greek, Latin, or Slavonic, or universal education. He created mathematical, navigation, artillery, and engineering schools for utilitarian purposes. In 1725 an Academy of

Sciences with a university and a *gimnaziya* (secondary school) was founded at St. Petersburg. The utilitarian, secular, and scientific characteristics of Peter's schools became the dominant features of Russian education, but, as a result of the many changes of policy after Peter's death in 1725, a national system of education did not develop.

A second attempt at nationalizing education in Russia was made by Catherine II. After many abortive schemes, Catherine issued in 1786 a statute for schools, which can be considered the first Russian education act for the whole country. According to this act, a two-year course in minor schools was to be started in every district town and a five-year course in major schools in every provincial town. Catherinian schools were also to be utilitarian, scientific, and secular. At the end of the 18th century, 254 towns had the new schools, but 250 smaller towns and the rural districts had no schools whatever.

A third nationalizing attempt was made by Alexander I and was influenced by the disintegration of the serf system, by the development of industry and commerce, and by the ideas of the French Revolution. The new statutes (1803 and 1804) maintained the principles of utility and secular scientific instruction. The parochial schools (*prikhodskiye uchilishcha*) in the rural areas were to instruct the peasantry in reading, writing, arithmetic, and elements of agriculture; the district schools of urban areas (*uyezdnye uchilishcha*) and the provincial schools (*gimnazii*) were to give instruction in subjects necessary for civil servants—law, political economy, technology, and commerce. The system was state-controlled and free and formed a continuous ladder to the universities. Later conservative reactions, however, tended to blunt or reverse these reforms.

England. In England the development of a "national" education took a completely different course. It was influenced not by a political but by an industrial revolution. It is true that theorists such as Adam Smith, Thomas Paine, and Thomas Robert Malthus proposed state organization of elementary-schooling, but even they wanted to see limited state influence; the state could pay the musicians but not call the tune. Not until 1802 did Parliament intervene in the development of education, when the Health and Morals of Apprentices Act required employers to educate apprentices in basic mathematics, writing, and reading. For the most part this remained only a demand, since the employers were not interested in such education.

The reluctance on the part of the state induced several philanthropists to form educational societies, principally for the education of the poor. In 1796, for example, the Society for Bettering the Conditions of the Poor was founded. A further impulse for elementary education stemmed from the Sunday schools, the first of which was founded in 1780 in Gloucester; by 1785 their numbers had so increased that the Sunday School Society was founded. The lessons in such schools, however, were mainly those of Bible reading.

The monitorial, or Lancasterian, system

The educators Andrew Bell and Joseph Lancaster played a major role in progress toward an elementary-school system. They realized that the root of the problem lay in the lack of teachers and in the lack of money to hire assistants. Therefore, first Bell developed, then Lancaster modified, the so-called monitorial system (also called the Lancasterian system), whereby a teacher used his pupils to teach one another. The use of children to teach other children was not new, but Bell and especially Lancaster took the approach and developed it into a systematic plan of education. From 200 to 1,000 children were gathered in one room and seated in rows, usually of 10 pupils each. An adult teacher taught the monitors, and then each monitor taught his row of pupils the lesson in reading, writing, arithmetic, spelling, or higher subjects. Besides monitors who taught, there were, in Lancaster's system, monitors to take attendance, give examinations, issue supplies, and so on; school activity was to be directed with military precision; the emphasis was on drill and memorization. The system and the publicity connected with it expanded the efforts toward mass education, even though, pedagogically, the whole process was so routinized and formalized that opportunities for creative thinking or initiative scarcely existed. (H.-J.I./J.J.Ch.)

EUROPEAN OFFSHOOTS IN THE NEW WORLD

Spanish and Portuguese America. With the Spanish conquerors of the New World, the conquistadores, came friars and priests who immediately settled down to educate the Indians and convert them. Because there was little separation of church and state, the Roman Catholic church assumed complete control of elementary education, and the early Franciscan and Dominican friars were followed by Augustinians, Jesuits, and Mercedarians.

Church control of colonial education

The first elementary school in the New World was organized in Mexico by the Franciscan Pedro de Gante in 1523 in Texcoco, followed in 1525 by a similar school in San Francisco. Because such schools in Mexico were designed for Indian children, the monks learned the native languages and taught reading, writing, simple arithmetic, singing, and the catechism. The schools of the *hospicio* of the bishop Vasco de Quiroga in Michoacán added agriculture, trades, and crafts to their curriculum.

Mestizo children, the issue of Spanish and Indian parents, were often abandoned; thus, special institutions appeared to collect and educate them—for example, the Girls' School and the School of San Juan de Letrán, founded by Viceroy Mendoza in New Spain, and the Bethlehemite schools of Guatemala and Mexico.

In the beginning, children of Spaniards born in the colonies, called Creoles, had tutors. Eventually, schools promoted by cabildos (municipal authorities) emerged.

During the 18th century the Enlightenment came to Latin America, and with it a more secular and widespread education. Among famous projects were those of Viceroy Vertiz y Salcedo in Argentina and two model schools, free for children of the poor, by Archbishop Francos y Monroy in Guatemala. In New Spain the College of the Vizcainas (1767) became the first all-girl lay institution.

Because of the social structure, riches and administrative privilege were held by an elite, the Creoles, and secondary education was specially organized to serve them. Originally, secondary schools existed only in the monasteries, but when the Jesuits arrived in the late 1560s they founded important *colegios* (secondary institutions) to prepare students who wanted to enter the universities. There also existed a few special *colegios* for the Indian nobility, such as the outstanding Santa Cruz de Tlaltelolco (1536) in Mexico and San Andres in Quito, both founded by the Franciscans for liberal arts studies. The Jesuits also established schools for the Indians, including El Príncipe (1619) in Lima and San Borja in Cuzco. All these schools were eventually closed because of the jealousy of the Spanish bureaucracy.

Though the Dominicans and Franciscans had been pioneers in education, the Jesuits became the most important teachers. They offered an efficient education, molded to contemporary requirements, in boarding schools, where the elite of the Spaniards born in the Americas studied. When their order was expelled in 1767, education was dealt a severe blow. In Portuguese Brazil, where the expulsion edict had been issued eight years earlier and where they had been the only educators, the royal chancellor was forced to make feeble attempts toward organizing a secular education. The Spanish king Charles III also took advantage of the occasion and founded some new institutions—the Academy of San Carlos, the School of Mining in Mexico, the Royal College of San Carlos in Buenos Aires—and modernized others.

The Latin-American universities

Traditionally, Spanish universities had been organized on the model either of Paris or of Bologna. The former was a *universitas magistrorum,* governed by professors organized in faculties, whereas the latter, as a *universitas scholarium,* received its corporate authority from the student body organized into "nations" that elected leaders to whom even the professors were subject. In 1551 the Council of the Indies authorized the founding of the first American universities, one in Mexico and one in Lima; academic government was placed in the hands of a *claustro,* or faculty, composed of the rector, the teachers, and the professors. Dedicated to general studies, the universities required a papal as well as a royal authorization.

The Royal Pontifical University of Mexico was the first to open its doors, in 1553. In the Spanish colonies even-

tually 10 major and 15 minor universities came into existence. The latter were actually colleges—nine Jesuit, four Dominican, one Franciscan, and one Augustinian—which, because they were located far from the closest university (minimally 200 miles), were given special authorization to grant higher degrees. In Brazil no university existed, and Portuguese born in the colony had to go to Portugal for study.

Though in Spain itself law reigned supreme, in the Americas theology became the principal chair. Teaching was in scholastic mode: it began with the reading of a classical text; then the professor explained the thesis or proposition and offered arguments pro and contra so that a conclusion in accord with Roman Catholic dogma would result.

(J.Z.V.)

French Québec. Soon after the founding of the Québec colony in 1608, the first organized educational activity began with missionary work among the Indians, carried on mainly by members of the Récollet and Jesuit orders and, from 1639, by Ursuline nuns. The first mission "school" recorded was that of Pacifique du Plessis, established in 1616 in Trois-Rivières (Three Rivers).

Christian efforts among the Indians were only a dimension of the religious purposes that framed educational activity in Old World France. Roman Catholic social philosophy allowed no compromise in the spiritual direction of education, and both in informal socialization patterns and in what formal provisions existed the doctrine and aim of religion coincided with that of education. At the general level, education was intended to produce religious conformation in thought and behaviour; at the higher level, education was to produce a progeny of clerical leadership. The paternalistic authority of church and monarch was carried from the Old to the New World, where it perhaps became even more pervasive, due to the initial absence of alternative institutional developments. In education, the exclusive role of the state (though not insignificant) was confined to financial subsidization. Authority for the institution of education was vested in the bishop of Québec.

Most of the nonreligious functions now associated with formal education were, in the 17th and 18th centuries, carried in other institutional sectors: the family, the community, the vocation. Just as there was no sharp break between church and school in formal learning, there was an easy transition between the information and behaviour necessary for work and life as transmitted in the course of various socialization experiences. Thus, the self-sustaining and isolated life of the farmers, the wild and solitary ways of the coureurs de bois (fur traders), the miniature of European manners and customs established in the cities by the gentry—all contained within their own cycle the educative procedures for life in that society. Education as a separate institution was understandably associated with learning not related to the business of life.

The schools of New France

Institutional forms found in French colonial Québec included parish schools, girls' schools, secondary schools, and vocational schools; and literacy records indicate that the provision for education was in sum comparable to that in the Old World. Parish or common schools were irregularly provided to afford the rudiments of literacy and religion. Because of the relative sparseness of educational resources, social classes were frequently mixed in these schools. Girls' schools were established in Québec City by the Ursulines from 1642 and by the Sisters of the Congrégation de Notre Dame from 1659, with a rudimentary curriculum but including a characteristic "finishing" of social graces appropriate to the French-Canadian girl. Vocational training was probably of least concern in this early period, but specific attempts to institutionalize this educational area were begun as early as 1668 with the establishment of the School for Arts and Trades in Saint Joachim, for instruction in agriculture and certain trades.

Secondary education was offered by the Jesuits from 1636. The Jesuit college, offering early training for eventual entrance into the priesthood, was conducted along characteristically Jesuit lines: militaristic discipline in conduct, unequivocal authority in method, classical curriculum in content. The classical curriculum pattern, comprising basically Latin, Greek, mathematics, philosophy, and theology, was to be essentially preserved in the French-Canadian development of *collèges classiques* for secondary education.

In 1663 Bishop Laval established in the city of Québec the *grand séminaire* as the apex of the educational "system," as the first French-Canadian "university." Shortly thereafter, he also established the preparatory *petit séminaire.*

Following the cession of Québec to Britain in 1763, education fell prey to political and cultural disruption. Although the British military and colonial government attempted to preserve the structure of French civil and religious institutions, the cultural integrity of the system was inevitably broken. Financial grants from France for education discontinued and were not replaced by the British government; recruitment to religious orders was restricted; and educational development was obstructed by the continual association of educational plans with cultural-religious controversies. The end of the 18th century saw French-Canadian education fall backward into neglect.

(R.F.L.)

British America. *New England.* The year 1630, chronicled in New England annals as the beginning of the Great Migration, witnessed the founding there of Puritanism as the established religion. Rejecting democracy and toleration as unscriptural, the Puritans put their trust in a theocracy of the elect that brooked no divergence from Puritan orthodoxy. So close was the relation between state and church that an offense against the one was an offense against the other and, in either case, "treason to the Lord Jesus." The early Puritans also put their confidence in centralized church governance; however, geographic reality forced them to settle for a localized, congregational administration, for impossible roads made land travel over any distance onerous and even dangerous, and thus the focal point of social and political life had to be the village. Small and constricted, a place where the vital necessities, sacred and profane, were within walking range of all and where one's conduct was exposed to constant public watch, the New England village was the prime mover of communal life.

Puritan education in New England

In Puritan moral theology the young, like the old, were sinners doomed by almost insurmountable odds to perdition. To God, indeed, even infants were depraved, unregenerate, and damned. Hence, the sooner the young learned the ground rules of the good society, as revealed in the Bible, the better. The task of teaching them first befell the parents. Later, when they were old enough, the burden was conferred upon the school. The first secondary school was probably the Boston Latin School. Founded in 1635, it was modeled on the grammar schools of England, which is to say that it put an overwhelming emphasis on the ancient languages and "humane learning and good literature." By the 1640s the idea of town-supported schooling had lost its novelty.

If towns braved the first steps in education, then the Commonwealth of Massachusetts did not trail far behind. In 1642 it ordered parents and masters of apprentices to see to it that their charges were instructed in reading, religion, and the colony's principal laws. Five years later, the General Court reinforced this enactment with yet another. Aimed at the "old deluder Satan," it undertook to thwart him from keeping "men from a knowledge of the Scriptures," by requiring every township of 50 households to commission someone to teach reading and writing. The law also directed towns of 100 families to furnish instruction in Latin grammar so that youth might be "fitted for the university." Finally, the measure required teachers to be paid by "parents or masters . . . or by the inhabitants in general." The measure was given only a pallid obedience, but its assumption that the state may compel the schooling of its young and that in order to support education it may impose taxes is pertinent to subsequent times.

The founding of Harvard College

The first colonists had scarcely settled when in 1636 the General Court appropriated £400 "towards a school or college." When two years later John Harvard died and left the institution his library and some £800, the grateful founders honoured their school with his name. Designed to train youth for important Puritan places, particularly in the ministry, the college accepted only those who could

read, write, and speak Latin in prose and verse, besides knowing Greek nouns and verbs familiarly. Once admitted, the student was lodged at the college, pledged to a blameless behaviour, and put upon a prescribed four-year course of grammar, rhetoric, logic, arithmetic, geometry, astronomy, ethics, ancient history, Greek, and Hebrew. If he weathered these hazards, he was made a bachelor of arts (B.A.), and, if ambition still roweled him, he could enroll for another three years to become a master of arts (M.A.).

So things sat until the century's passing. Then, swayed by the intellectual breezes of Europe's Enlightenment, Harvard College ventured some earnest renovation. Its texts, cobwebbed with Aristotelianism, were replaced with newer ones by Locke and Sir Isaac Newton. In 1718 it added mathematics and sciences to its offerings, and 20 years later it enriched itself with a professorship of mathematics and natural philosophy. There were the usual grumblings from conservatives, and in 1701 a number of Congregational parsons, all Harvard sons, distressed by their alma mater's dalliance in newfangled ideas, inaugurated the collegiate school of Connecticut, now Yale University.

The new academies. Disdainful of the challenging intellectual values, the secondary schools continued in their classical tracks. By the 18th century, however, their tradition was playing out, especially among the rising nabobs of the marketplace. When the old schools failed to respond to their demands for an education calculated to prepare their sons for everyday living, they resorted to private schooling. From such endeavour emerged the academy. The first school of strictly native provenance, it made its advent in 1751 in Philadelphia (the Philadelphia Academy), the work in the main of Benjamin Franklin. What differentiated it from its classical antecedent was its promotion of "useful learning," to wit, the vernacular, modern languages, history, geography, chronology, navigation, mathematics, natural and applied science, and the like.

The first academies addressed themselves solely to boys, but time saw them vouchsafe instruction to girls in a "female department," which in turn gave way to the "female academy," whose curriculum reflected debates of the time about female education. Fine arts, domestic subjects, and training for occupations open to women were included, though some female educators stressed intellectual attainment rather than practical learning.

Private ventures always, academies generally were not loath to solicit outside assistance—some, indeed, as in New York, enjoyed a public subsidy. Whatever their special character, to their very end they maintained their original purpose of bringing education into closer consonance with "the great and the real business of living," as Phillips Academy of Andover, Massachusetts, phrased it when, in 1778, it held its first sessions.

Diversity of the middle colonies

The middle colonies. The religious uniformity that marked the Puritan theocracy was missing in the middle colonies. From New York through Delaware there flourished a host of sects whose scriptural interpretations were diverse—often, in fact, in collision. Nor was there even the tie of a common language, for the settlers came from many lands. Divergent in religion and language, the bedrock in those times of elementary schooling, the middle colonists could not accommodate themselves, as did the Puritans, to a single school teaching reading and religion to all the children of the neighbourhood. Instead, they depended on parish or parochial schools, each of them free to teach by its own denominational lights. True, for a time New Netherland, with its established Dutch Reformed church, maintained some town schools, but, after the English seized the colony (renaming it New York), such endeavours ceased. Pennsylvania, linguistically and denominationally the most heterogeneous of the colonies, began its educational history by ordering the erection of public schools and the instruction of children. But the ordinance fell prey to powerful sectarian antagonisms, and in 1701 the colony essayed to make peace by sanctioning the establishment of parochial schools.

Like the New Englanders, the middle colonists aspired to establish colleges, but, with no friendly lawmakers to sustain them, they found their task heavily hobbled, and the mid-1700s were upon them before their hopes materialized with the advent, in 1746, of the College of New Jersey (Princeton). There followed King's College (Columbia) in 1754; the College and Academy of Philadelphia (Pennsylvania) in 1755; and Queen's College (Rutgers) in 1766. Common to these schools was their stress on the ancient languages, metaphysics, and divine science. At the same time, however, one discerns signs of a new liberalism. Both Rutgers and Columbia announced their interdenominationalism. Pennsylvania offered courses in physics, and in 1765 it became the first colonial college to sponsor systematic instruction in medicine.

Effects of plantation life on Southern education

The Southern colonies. Unlike New Englanders, Southerners resided not in villages but on widely scattered plantations. For years, town life was impossible and so, per consequence, were town schools. But even had their establishment been feasible, the odds against them were staggering, since the ruling classes, like their analogues overseas in England, were averse to schooling the young under governmental direction. Instead, they regarded education as a personal concern, the affair of parent and church rather than of the state. Left thus to their own devices, Southerners schooled their young to suit their taste, the rich resorting to tutors and private schools and the rest scratching out an education as best they could. Time saw the appearance of a number of free schools serving those who were neither rich nor poor. For the offspring of the low-down and unregarded folk, Virginia enacted its law of 1642. An echo of England's Poor Law, it provided for the "relief of such parents whose poverty extends not to give them [the children] breeding." For this purpose it ordered the creation of a "workhouse school" at James City to which each county was to commit two children of an age of six or over. There, besides being reared as Anglicans, they were to be "instructed in honest and profitable trades and manufactures as also to avoid sloth and idleness." Amended several times, the statute became the model for similar legislation throughout the South.

The first Southern college was founded in Virginia in 1693. William and Mary College was chartered to propagate the "Liberal Arts and the Christian Faith," with particular stress on preparing young men for the Anglican pulpit. As the 18th century swept on, the secular interest that had invaded Harvard appeared in Virginia, and there ensued a waning of the earlier religious motivation. In 1779, led by Thomas Jefferson, the college trustees refurbished the school with chairs in medicine, mathematics, physics, moral philosophy, economics, law, and politics. The chair in divinity was discontinued as "incompatible with freedom in a republic." (A.E.M./R.F.L.)

Newfoundland and the Maritime Provinces. Newfoundland was, during most of this period, under British control, and, though there were settlers even before the 17th century, the island was not considered a settlement colony. Other than for naval training and fishing advantages, the British government had no concern for Newfoundland. Thus, policies were constructed with regard to the rights and advantages of British seamen, while, implicitly as well as in overt regulations, settlement was obstructed and restricted. Destruction from the running French-British military conflicts further discouraged development. These conditions of economic and political diminution of the settlement from outside were aggravated by the usurious conduct of merchants and the corruption of officials and by the national and religious divisions among the inhabitants themselves.

With such substantial problems of mere survival in Newfoundland, it is not surprising that the luxury of formal education was almost absent during this period. Some accounts verify that informal, unorganized efforts were made on an occasional basis to convey minimum schooling to settlers' children, but the only organized effort was that of the Society for the Propagation of the Gospel in Foreign Parts (SPGFP). The SPGFP founded or aided a school in Bonavista in 1722 and in St. John's in 1744 and sponsored schools in more than 20 settlements between 1766 and 1824. Religion was undoubtedly more important than education as such to the society, but its provision of reading materials as well as the mere act of establishing some kind of school filled a notable void in the Newfoundland

Society for the Propagation of the Gospel in Foreign Parts

settlement. Other charitable societies, such as the Society for Improving the Condition of the Poor in St. John's, the Benevolent Irish Society, the Newfoundland School Society (later the Colonial and Continental Church Society), the Wesleyan Society, the Sisters of the Presentation, the Sisters of Mercy, and the Irish Christian Brothers, carried the charity-school work into the 19th century and maintained a thread of education through the colonial "dark ages."

For a time after 1763 the Maritimes were all one colony—Nova Scotia—but Prince Edward Island was separated in 1769 and New Brunswick in 1784. This area comprised a heterogeneous population of French Acadians, English Protestants and others from Europe, Highland Scots, and loyalists from the United States. Each of these groups carried attitudes more or less favourable to education, and the regionalization of these attitudes, together with other conditions, influenced the differential development of education in the area. At the end of the 18th century, for example, New Brunswick, with a high loyalist population promoting political and educational development, probably ranked highest among the Maritime colonies in educational interest.

The first relatively organized attempt at common schooling in the Maritimes was made by the SPGFP, closely connected to the Church of England. The society opened both weekday and Sunday schools, and it might be said that it fostered teacher training in stipulating qualifications for its teachers. Other than SPGFP schools, education in the Maritime colonies was carried on by itinerant teachers and in scattered private-venture schools. Schools for separate ethnic or religious groups were discouraged by the Anglicans, but consistent pressure for such schools did succeed, at least temporarily; for example, in Lunenburg, Nova Scotia, and in Sydney, on Cape Breton Island. A school for blacks was established in Halifax in 1788.

Upper schools were established only toward the end of the 18th century in the Maritimes. As they were established singularly and recruited from a social class rather than from a lower school, there is no clear line of demarcation among the various types as there would be later in an integrated system. Basically, they were Anglican and classical, although the private schools, advertising to as wide a clientele as possible, often included some breadth, extending into practical studies. Probably the most influential of the early attempts were the two Latin grammar schools founded in 1788 and 1789 at Windsor and Halifax, Nova Scotia. The former became associated with King's College, established in Windsor at the same time. Thomas McCulloch's Academy at Pictou, Nova Scotia, and the College of New Brunswick at Fredericton, both founded around the turn of the century, were also early exemplars of higher education. (R.F.L.)

Western education in the 19th century

THE SOCIAL AND HISTORICAL SETTING

From the mid-17th century to the closing years of the 18th century, new social, economic, and intellectual forces steadily quickened—forces that in the late 18th and the 19th centuries would weaken and, in many cases, end the old aristocratic absolutism. The European expansion to new worlds overseas had stimulated commercial rivalry. The new trade had increased national wealth and encouraged a sharp rise in the numbers and influence of the middle classes. These social and economic transformations, joined with technological changes involving the steam engine and the factory system, together produced industrialism, urbanization, and the beginnings of mass labour. At the same time, intellectuals and philosophers were assaulting economic abuses, old unjust privileges, misgovernment, and intolerance. Their ideas, which carried a new emphasis on the worth of the individual—the citizen rather than the subject—helped not only to inspire political revolutions, sometimes successful, sometimes unsuccessful, but, more important, to make it impossible for any government, even the most reactionary, to disregard for long the welfare of the common man. Finally, there was a widespread psychological change: man's confidence in his power to use resources, master nature, and structure his own future was heightened beyond anything known before; and this confidence on a national scale—in the form of nationalism—moved all groups to struggle for the freedom to direct their own affairs.

Education as the responsibility of the state

All these trends influenced the progress of education. One of the most significant results was the gradual acceptance of the view that education ought to be the responsibility of the state. Some countries, such as France and Germany, were inspired by a mixture of national aspiration and ideology to begin the establishment of public educational systems early in the 19th century. Others, such as Great Britain and the United States, under the spell of laissez-faire, hesitated longer before allowing the government to intervene in educational affairs. The school reformers in these countries had to combat the prevailing notion that "free schools" were to be provided only for pauper children, if at all; and they had to convince society that general taxation upon the whole community was the only adequate way to provide education for all the children of all the people.

The new social and economic changes also called upon the schools, public and private, to broaden their aims and curricula. Schools were expected not only to promote literacy, mental discipline, and good moral character but also to help prepare children for citizenship, for jobs, and for individual development and success. Although teaching methods remained oriented toward textbook memorizing and strict discipline, a more sympathetic attitude toward children began to appear. As the numbers of pupils grew rapidly, individual methods of "hearing recitations" by children began to give way to group methods. The monitorial, or Lancasterian, system became popular because, in the effort to overcome the shortage of teachers during the quick expansion of education, it enabled one teacher to use older children to act as monitors in teaching specific lessons to younger children in groups. Similarly, the practice of dividing children into grades or classes according to their ages—a practice that began in 18th-century Germany—was to spread everywhere as schools grew larger.

THE EARLY REFORM MOVEMENT: THE NEW EDUCATIONAL PHILOSOPHERS

Influence of Rousseau

The late 18th and 19th centuries represent a period of great activity in reformulating educational principles, and there was a ferment of new ideas, some of which in time wrought a transformation in school and classroom. The influence of Rousseau was profound and inestimable. One of his most famous followers was Pestalozzi, who believed that children's nature, rather than the structure of the arts and sciences, should be the starting point of education. Rousseauist ideas are seen also in the work of Friedrich Froebel, who emphasized self-activity as the central feature of childhood education, and in that of Johann Friedrich Herbart, perhaps the most influential 19th-century thinker in the development of pedagogy as a science.

Pestalozzi. The theories of the Swiss reformer Johann Heinrich Pestalozzi laid much of the foundation of modern elementary education. Beginning as a champion of the underprivileged, he established near Zürich in 1774 an orphanage in which he attempted to teach neglected children the rudiments of agriculture and simple trades in order that they might lead productive, self-reliant lives. A few years later the enterprise failed, and Pestalozzi turned to writing, producing his chief work on method, *How Gertrude Teaches Her Children,* in 1801, and then began teaching again. Finally in 1805 he founded at Yverdon his famous boarding school, which flourished for 20 years, was attended by students from every country in Europe, and was visited by many important figures of the time, including the philosopher Johann Gottlieb Fichte, the educators Froebel and Herbart, and the geographer Carl Ritter.

The pedagogy of Pestalozzi. In spite of the quantity of his writings, it cannot be said that Pestalozzi ever wrote a complete and systematic account of his principles and methods; an outline of his theories must be deduced from his various writings and his work. The foundation of his doctrine was that education should be organic, meaning that intellectual, moral, and physical education (or, in his

Child-centred education

words, development of "head, heart, and body") should be integrated and that education should draw upon the faculties or "self-power" inherent in the human being. Education should be literally a drawing-out of this self-power, a development of abilities through activity—in the physical field by encouraging manual work and exercises, in the moral field by stimulating the habit of moral actions, and in the intellectual field by eliciting the correct use of the senses in observing concrete things accurately and making judgments upon them. Words, ideas, practices, and morals have meaning only when related to concrete things.

From these overarching principles there followed certain practical rules of educational method. First, experience must precede symbolism. There must be an emphasis on object lessons that acquaint the child with the realities of life; from these lessons abstract thought is developed. What one does is a means to what one knows. This means that the program should be child-centred, not subject-centred. The teacher is to offer help by participating with the child in his activities and should strive to know the nature of the child in order to determine the details of his education. This means that the stages of education must be related to the stages of child development. Finally, intellectual, moral, and physical activities should be as one.

Much of Pestalozzi's pedagogy was influenced by his work with children of the poor. Thus, there was a strong emphasis on education in the home. The development of skills was emphasized, not for their own sake, but in connection with intellectual and moral growth. Manual training was important for the head and heart, as well as for the hand. Whereas the reformers of the Enlightenment and the French Revolution stressed the "emancipation" of the lower classes, Pestalozzi aimed at helping poor people to help themselves. This was social reform, not social revolution.

The influence of Pestalozzi. "The art of education," Pestalozzi claimed, "must be significantly raised in all its facets to become a science that is to be built on and proceeds from the deepest knowledge of human nature." By his own efforts in this direction, he stimulated pedagogical theory and practice to an enormous degree in many parts of the Western world. By his philanthropic efforts on behalf of the poor, he inspired new movements toward the reform of philanthropic educational institutions and the pedagogy applied to such institutions; he created a new methodology for elementary education that was introduced not only into schools but also into programs of teacher education in Europe and America; and by his own example he gave teachers a high professional ethos. Pestalozzi, like few others at any time, recognized and sincerely tried to alter the misery existing in the world. If the Enlightenment saw its pedagogical mission as the spreading of the light of reason, then Pestalozzi showed that it was not reason alone but love above all that would show a way out of the "mire of the world."

Pestalozzi's influence abroad

It is hardly possible to name all of Pestalozzi's disciples—the Pestalozzians—for almost all the pedagogical figures of his time literally or figuratively went to his school. His influence was most profound in Germany, especially in Prussia and Saxony. Generally speaking, in the first half of the 19th century the English school system was completely under the influence of the disciplinarian monitorial systems of Bell and Lancaster. Pestalozzi for most Englishmen was "a distressing type of the German" and "an idealistic dreamer," as some critics put it. Nevertheless, he exercised some influence in England through James Pierrepont Greaves and the London Infant School Society and through Charles and Elizabeth Mayo and the Home and Colonial School Society. In the United States Pestalozzianism was introduced by a Philadelphia scientist and philanthropist, William Maclure, one of the sponsors of the utopian colony at New Harmony, Ind., and by Joseph Neef, who opened a school near Philadelphia.

In Switzerland itself, in Hofwil near Bern, Philipp Emanuel von Fellenberg founded an institution for the education of the poor. He tried to build up a kind of pedagogical province or miniature state, in which work was the means of self-help and in which the pedagogical program was the joint responsibility of teachers and pupils.

Froebel and the kindergarten movement. Next to Pestalozzi, perhaps the most gifted of early 19th-century educators was Froebel, the founder of the kindergarten movement and a theorist on the importance of constructive play and self-activity in early childhood. He was an intensely religious man who tended toward pantheism and has been called a nature mystic. Throughout his life he achieved very little literary fame, partly because of the style of his prose and philosophy, which is so academic and obscure that it is difficult to read and sometimes scarcely comprehensible.

In early life, Froebel tried various kinds of employment until in 1805 he met Anton Gruner, a disciple of Pestalozzi and director of the normal school at Frankfurt am Main, who persuaded him to become a teacher. After two years with Gruner, he visited Pestalozzi at Yverdon, studied at Göttingen and Berlin, and eventually determined upon establishing his own school, founded on what he considered to be psychological bases. The result in 1816 was the Universal German Educational Institute at Griesheim, transferred the following year to Keilhau, which constituted a kind of educational community for Froebel, his friends, and their wives and children. To this period belongs *The Education of Man* (1826), his most important treatise, though typical of his obscurantism. In 1831 he was again in Switzerland, where he opened a school, an orphanage, and a teacher-training course. Finally, in 1837, upon returning to Keilhau, he opened his first *Kindergarten,* or "garden of children," in nearby Bad Blankenburg. The experiment attracted wide interest, and other kindergartens were started and flourished, despite some political opposition.

The first kindergarten

The pedagogy of Froebel. Froebel's pedagogical ideas have a mystical and metaphysical context. He viewed man as a child of God, of nature, and of humanity who must learn to understand his own unity, diversity, and individuality, corresponding to this threefold aspect of his being. On the other hand, man must understand the unity of all things (the pantheistic element).

> Education consists of leading man, as a thinking, intelligent being, growing into self-consciousness, to a pure and unsullied, conscious and free representation of the inner law of Divine Unity, and in teaching him ways and means thereto.

Education had two aspects: the teacher was to remove hindrances to the self-development or "self-activity" of the child, but he was also to correct deviations from what man's experience has taught is right and best. Education is thus both "dictating and giving way." This means that ordinarily a teacher should not intervene and impose mandatory education, but when a child, particularly a child of kindergarten age, is restless, tearful, or willful, the teacher must seek the underlying reason and try to eradicate the uncovered hindrance to the child's creative development. Most important, the teacher's dictating and giving way should not flow from the mood and caprices of the teacher. Behaviour should be measured according to a "third force" between teacher and child, a Christian idea of goodness and truth.

School, for Froebel, was not an "establishment for the acquisition of a greater or lesser variety of external knowledge"; actually, he thought children were instructed in things they do not need. School instead should be the place to which the pupil comes to know the "inner relationship of things," "things" meaning God, man, nature, and their unity. The subjects followed from this: religion, language and art, natural history, and the knowledge of form. In all these subjects the lessons should appeal to the pupil's interests. It is clear that, in Froebel's view, the school is to concern itself not primarily with the transmission of knowledge but with the development of character and the provision of the right motivation to learn.

Froebel's emphasis on play

Froebel put great emphasis on play in child education. Just like work and lessons, games or play should serve to realize the child's inner destiny. Games are not idle time wasting; they are "the most important step in the development of a child," and they are to be watched by the teachers as clues to how the child is developing. Froebel was especially interested in the development of toys for children—what he called "gifts," devised to stim-

ulate learning through well-directed play, These gifts, or playthings, included balls, globes, dice, cylinders, collapsible dice, shapes of wood to be put together, paper to be folded, strips of paper, rods, beads, and buttons. The aim was to develop elemental judgment distinguishing form, colour, separation and association, grouping, matching, and so on. When, through the teacher's guidance, the gifts are properly experienced, they connect the natural inner unity of the child to the unity of all things: *e.g.*, the sphere gives the child a sense of unlimited continuity, the cylinder a sense both of continuity and of limitation. Even the practice of sitting in a circle symbolizes the way in which each individual, while a unity in itself, is a living part of a larger unity. The child is to feel that his nature is actually joined with the larger nature of things.

The kindergarten movement. The kindergarten was unique for its time. Whereas the first institutions for small children that earlier appeared in Holland, Germany, and England had been welfare nursery schools or day-care centres intended merely for looking after children while parents worked, Froebel stood for the socializing or educational idea of providing, as he put it in founding his kindergarten, "a school for the psychological training of little children by means of play and occupations." The school, that is, was to have a purpose for the children, not the adults. The curriculum consisted chiefly of three types of activities: (1) playing with the "gifts," or toys, and engaging in other occupations designed to familiarize children with inanimate things, (2) playing games and singing songs for the purpose not only of exercising the limbs and voice but also of instilling a spirit of humanity and nature, and (3) gardening and caring for animals in order to induce sympathy for plants and animals. All this was to be systematic activity.

The kindergarten plan to meet the educational needs of children between the ages of four and six or seven through the agency of play thereafter gained widespread acceptance. During the 25 years after Froebel's death in 1852, kindergartens were established in leading cities of Austria, Belgium, Canada, Germany, Great Britain, The Netherlands, Hungary, Japan, Switzerland, and the United States. In Great Britain the term infant school was retained for the kindergarten plan, and in some other countries the term crèche has been used.

Widespread acceptance of kindergartens

Herbart. Johann Friedrich Herbart was a contemporary of Froebel and other German Romanticists, but he can hardly be put into the ranks of such pedagogues. During his lifetime his sober, systematic "philosophical realism" found little approval; and only posthumously, during the latter half of the 19th century, did his work achieve great importance. He is regarded as one of the founders of theoretical pedagogy, injecting both metaphysics and psychology into the study of how people learn.

The psychology and pedagogy of Herbart. As a young man of 18, Herbart had studied at the University of Jena under the idealist philosopher Fichte. It was a long while before he broke from the spell of Fichte's teachings and turned to philosophical realism, which asserts that underlying the world of appearances there is a plurality of things or "reals." Change consists simply in the alteration in the relations between these reals, which resist the changed relationships as a matter of self-preservation.

Ideas, like things, always exist and always resist change and seek self-preservation. It is true that some ideas may be driven below the threshold of consciousness; but the excluded ideas continue to exist in an unconscious form and tend, on the removal of obstacles (as through education), to return spontaneously to consciousness. In the consciousness there are ideas attracting other ideas so as to form complex systems. These idea masses correspond to the many interests of the individual (such as his home and his hobbies) and to broader philosophical and religious concepts and values. In the course of mental development certain constellations of ideas acquire a permanent dominance that exercises a powerful selective facilitating influence upon the ideas struggling to enter or reenter the consciousness.

In his systematic account of the nature of education, Herbart conceived the process as beginning with the idea masses that the child has previously acquired from experience and from social intercourse. The teacher creates knowledge from the former and sympathy from the latter. The ultimate objective is the formation of character by the development of an enlightened will, capable of making judgments of right and wrong. Moral judgments (like reals) are absolute, springing from contemplation, incapable of proof and not requiring proof. Ethics, in other words, is the ultimate focus of pedagogy.

In the classroom, it is the aim of the lessons to introduce new conceptions, to bind them together, and to order them. Herbart speaks of "articulation," a systematic method of constructing correct, or moral, idea masses in the student's mind. First the student becomes involved in a particular problem; then he considers its context. Each of these two stages has a phase of rest and of progress, and thus there are four stages of articulation: (1) clarification, or the static contemplation of particular conceptions, (2) association, or the dynamic linking of new conceptions with old ones, (3) systematization, or the static ordering and modification of what in the conceptions are deemed of value, and (4) methodization, or the dynamic application and recognition of what has been learned. Herbart phrased this system of instruction only in very general terms, but his successors tended to turn this framework into a rigid schedule that had to be applied to every lesson. Herbart himself warned:

Articulation, or the systematization of lessons

> We must be familiar with them [the methods], try them out according to circumstances, alter, find new ones, and extemporize; only we must not be swallowed up in them nor seek the salvation of education there.

The Herbartians. Herbart's basing of educational methods on an understanding of mental processes or psychological considerations, his view that psychology and moral philosophy are linked, and his idea that instruction is the means to moral judgment had a large place in late 19th-century pedagogical thought. Among Herbart's followers were Tuiskon Ziller in Leipzig (who was the founder of the Association for Scientific Pedagogy) and Wilhelm Rein in Jena. From 1895 to 1901 a National Herbart Society for the Scientific Study of Education flourished in the United States; John Dewey was a major critic of Herbartianism in its proceedings.

Ziller's ideas are representative of the Herbartians. He insisted that all parts of the curriculum be closely integrated and unified—history and religion forming the core subjects on which everything else was hinged. The sequence of instruction was to be adjusted to the psychological development of the individual, which was seen as corresponding to the cultural evolution of mankind in stages from primitive savagery to civilization. His main aim in education, like the aim of most Herbartians, was promoting character building, not simply knowledge accumulation.

Other German theorists. In the history of pedagogy there is no period of such fruitfulness as the 19th century in Germany. In addition to Herbart, Froebel, Pestalozzi (in German Switzerland), and their followers, there were scores of the most important writers, philosophers, and theologians contributing their ideas on education—including Friedrich Schiller, Johann Wolfgang von Goethe, G.W.F. Hegel, Friedrich Ludwig Jahn, Johann Paul Friedrich Richter, Ernst Moritz Arndt, and Friedrich Nietzsche. To list the many ideas and contributions of these figures and others is impossible here, but it is worthwhile to suggest briefly the work of three men—Johann Gottlieb Fichte, Friedrich Schleiermacher, and Wilhelm von Humboldt—representing three divergent views.

When the great heterodox University of Berlin was founded in 1809, Fichte became one of its foremost professors and a year later its second rector, having already achieved fame throughout Germany as an idealist philosopher and fervent nationalist. At a time when Napoleon had humbled Prussia, Fichte in Berlin delivered the powerful *Addresses to the German Nation* (1807–08), full of practical views on national recovery and glory, including suggestions on the complete reorganization of the German schools along Pestalozzian lines. All children would be educated—and would be educated by the state. Boys and girls would be taught together, receiving virtually the same

Fichte's ordered education

education. There would be manual training in agriculture and the industrial arts, physical training, and mental training, the aim of which would be not simply the transmission of measures of knowledge but rather the instillation of intellectual curiosity and love and charity toward all men. Unlike Pestalozzi, however, Fichte was wary of the influence of parents and preferred educating children in a "separate and independent community," at least until a new generation of parents had arisen, educated in the new ideas and ideals. Here was an apparent revival of Plato's idea of a strictly ordered, authoritarian state.

Another of the founders of the University of Berlin (teaching there from 1810 to 1834) was the Protestant theologian Friedrich Schleiermacher, who sounded a very modern note by offering a social interpretation of education. Education, in his view, was an effort on the part of the older generation to "deliver" the younger generation into the four spheres of life—church, state, social life, and science. Education, however, not only assumes its organization in terms of these four areas of life but also serves to develop and influence these areas.

Perhaps more than any other individual, the philologist and diplomat Wilhelm von Humboldt was responsible for the founding of the University of Berlin. Supported by the king of Prussia, Frederick William III, he adopted for it principles that raised it to a foremost place among the universities of the world—the most important principle being that no teacher or student need adhere to any particular creed or school of thought. This academic freedom survived in Germany despite its temporary suspension and Humboldt's dismissal by a reactionary Prussian government in 1819. Philosophically and pedagogically, Humboldt was himself a humanist—a part of a wave of what were called new humanists—who reasserted the importance of studying the classical achievements of humanity in language, literature, philosophy, and history. The aim of education in these terms was not the service of society or the state but rather the cultivation of the individual.

Humboldt's humanism

French theorists. At this time there were two men in France who were making their names through the introduction of new methods—Jean-Joseph Jacotot and Édouard Séguin. Jacotot was a high-school teacher, politician, and pedagogue, whose main educational interests focused on the teaching of foreign languages. "You learn a foreign language," he said, "as you learn your mother-language." The pupil is confronted with a foreign language; he learns a text in the language almost by heart, compares it with a text in his own native language, and then tries gradually to free himself from the comparison of texts and to construct new combinations of words. The teacher controls this learning by asking questions. "My method is to learn one book and relate all the others to it." The learning of grammar came later.

Jacotot's method emphasized first the practical side and then the rule, constant repetition, and self-activity on the part of the pupils. Controversy arose, however, over his two basic theses: (1) that everyone has the same intelligence, differences in learning success being only a case of differences in industry and stamina, and (2) that everything is in everything: "Tout est dans tout," which suggests that any subject or book is analogous to any other.

The doctor and psychologist Édouard Séguin developed a pedagogy for pupils of below-average intelligence. Historically, scientific attempts to educate mentally retarded children had begun with the efforts of a French doctor, Jean-Marc-Gaspard Itard, during the latter part of the 18th century. In his classic book, *The Wild Boy of Aveyron* (1801), Itard related his five-year effort to train and educate a boy found, at about the age of 11, running naked and wild in the woods of Aveyron. Later, Séguin, a student of Itard, devised an educational method using physical and sensory activities to develop mental processes. Limbs and the senses were, in his view, a part of the whole personality, and their development was a part of the whole human education. His method was a specific adaptation of the idea that the development of intellectual and moral distinctions grows out of sensory experience.

Education of the mentally retarded

Spencer's scientism. The English sociologist Herbert Spencer was perhaps the most important popularizer of science and philosophy in the 19th century. Presenting a theory of evolution prior to Charles Darwin's *On the Origin of Species by Means of Natural Selection,* Spencer argued that all of life, including education, should take its essential lessons from the findings of the sciences. In *Education: Intellectual, Moral, and Physical* (1860) he insisted that the answer to the question "What knowledge is of most worth?" is the knowledge that the study of science provides. While the educational methodology Spencer advocated was a version of the sense realism espoused by reformers from Ratke and Comenius down to Pestalozzi, Spencer himself was a social conservative. For him, the value of science lies not in its possibilities for making a better world but in the ways science teaches man to adjust to an environment that is not susceptible to human engineering. Spencer's advocacy of the study of science was an inspiration to the American Edward Livingston Youmans and others who argued that a scientific education could provide a culture for modern times superior to that of classical education. (H.-J.I./J.J.Ch.)

DEVELOPMENT OF NATIONAL SYSTEMS OF EDUCATION

The great changes in Europe in the 19th century included, among other things, the further consolidation of national states, the spread of modern technology and industrialization, and increasing secularization. These changes had consequences for the design of school systems. National school systems had to be conceived and organized. Alongside the older schools, including elementary schools, Latin, or grammar, secondary schools, and universities, there developed so-called modern schools that stressed the exact sciences and modern languages, reflecting the new technological and commercial age. Vocational schools also appeared in greater numbers. The influence of the church was increasingly repressed, and the influence of the state on the school system correspondingly grew stronger. The ideal of universal education—education for all—became more and more a reality.

Germany. Luther's pronouncements on the educational responsibilities of the individual had no doubt helped create that healthy public opinion that rendered the principle of compulsory school attendance acceptable in Prussia at a much earlier date than elsewhere. State intervention in education was almost coincident with the rise of the Prussian state. In 1717 Frederick William I ordered all children to attend school if schools were available to them. This was followed in 1736 by edicts for the establishment of schools in certain provinces, in 1763 by Frederick II the Great's regulation asserting the principle of compulsory school attendance, and in 1794 by a codification of Prussian law recognizing the principle of state supremacy in education.

Humboldt's reforms. The schools, however, had established a traditional classical curriculum that ignored the changing needs of life and fields of knowledge. No effective reorganization of the educational system was carried out until after the disaster of the Battle of Jena (1806), during the Napoleonic Wars, which brought about the virtual collapse of Prussia. Fichte delivered his *Addresses to the German Nation* at this time, appealing to the spirit of patriotism over a selfish individualism. He advocated a nationalism to be cultivated and enhanced by controlling the education of the young. In the period of governmental reform which came about, one of the first acts of the prime minister Freiherr Karl vom Stein in 1807 was to abolish certain semi-ecclesiastical schools and to place education under the Ministry of the Interior, with Wilhelm von Humboldt at the head of a special section. Humboldt's policy in secondary education was a compromise between the narrow philological pedantry of the old Latin schools and the large demands of the new humanism that he espoused. The measure introduced by Humboldt in 1810 for the state examination and certification of teachers checked the then-common practice of permitting unqualified theological students to teach in the schools and raised the teaching profession to a high level of dignity and efficiency, placing Prussia in the forefront of educational progress. It was also a result of the initiative of Humboldt that the methods of Pestalozzi were

Reforms inspired by the Napoleonic disasters

introduced into the teachers' seminaries. To this period also belongs the revival, in 1812, of the *Abitur* (the school-leaving examination), which had fallen into abeyance.

Developments after 1815. The period that succeeded the peace of 1815 was one of political reaction, and not until the 1830s were there further significant reforms. In 1834, for example, an important step was taken in regard to secondary education by making it necessary for candidates for the learned professions, as well as for the civil service and for university studies, to pass the leaving examination of the *Gymnasien,* the classical secondary schools. Thus, through the leaving examination, the state held the key to the liberal careers and was thereby able to impose its own standards upon all secondary schools.

In connection with the so-called Kulturkampf, or struggle between the state and the Roman Catholic church, the school law of 1872 reasserted the absolute right of the state alone to the supervision of the schools. Nevertheless, the Prussian system remained both for Catholics and for Protestants essentially denominational. On the elementary level, in particular, the mixed school was established only when the creeds were so intermingled that a confessional school was impracticable. In all cases the teachers were appointed with reference to religious faith; religious instruction was given in school hours and was inspected by the clergy.

Types of German secondary schools

The official classification, or grading according to the type of curriculum, of secondary schools in Prussia (and throughout Germany) was very precise. The following were the officially recognized types: (1) the classical nine-year *Gymnasium,* with a curriculum that included Latin, Greek, and a modern language, (2) the semiclassical nine-year *Realgymnasium,* with a more modern curriculum that included, in addition to Latin and modern languages, the natural sciences and mathematics, and (3) the modern six-year *Realschule* or nine-year *Oberrealschule,* with a curriculum of sciences and mathematics.

The differentiation between the types was the result of a natural educational development corresponding to the economic changes that transformed Prussia from an agricultural to an industrial state. The classical schools long retained their social prestige and a definite educational advantage in that only their pupils were admissible to the universities. From the foundation of the German Empire in 1871 the history of secondary education was largely concerned with a struggle for a wider recognition of the work of the newer schools. The movement received a considerable impetus by the action of Emperor William II, who summoned a school conference in 1890 at which he set the keynote: "It is our duty to educate young men to become young Germans and not young Greeks or Romans." New schedules were framed in which the hours devoted to Latin were considerably reduced, and no pupil could obtain a leaving certificate without a satisfactory mark in the mother tongue. The reform lasted only a single school generation. In 1900 equality of privileges was granted to three types of schools, subject to certain reservations: the theological faculties continued to admit only students from classical schools, and the pupils of the *Oberrealschule* were excluded by their lack of Latin from the medical faculties; but insofar as Latin was required for other studies, such as law or history, it could be acquired at the university itself.

Girls' schools. In Prussia, as elsewhere, the higher education of girls lagged far behind that of boys and received little attention from the state or municipality, except insofar as the services of women teachers were needed in the elementary schools. Thus it came about that in Prussia secondary schools for girls were dealt with administratively as part of the elementary-school system. After the establishment of the German Empire in 1871, a conference of directors and teachers of these schools was held at Weimar and put forth a reasoned plea for better organization and improved status. The advocates of reform, however, were not at unity in their aims; some wished to lay stress on ethical, literary, and aesthetic training; others stressed intellectual development and claimed an equal share in all the culture of the age. Even the women teachers fought an unequal battle, for all the school heads and a large part of the staff were men, usually academically trained. The women continually demanded a larger share of the work, and this was secured by the establishment of a new higher examination for women teachers. University study, though not prescribed, was in fact essential, and yet women had not the right of access to the university in Germany. They were allowed to take the leaving examination, for which private institutions prepared them, but their admission to the university rested with the professor. Not until the 20th century were desired changes achieved.

The new German universities. Unquestionably one of the greatest worldwide influences exercised by German education in the 19th century was through its universities, to which students came from all over the world and from which every land drew ideas for the reformation of higher education. To understand this, one must be aware of the state of higher education in most countries in the 19th century. Although the century witnessed a steady expansion of scientific knowledge, the curriculum of the established universities went virtually untouched. Higher education followed a single dimension. This was the century of the scientists Michael Faraday, Hermann von Helmholtz, James Prescott Joule, Charles Darwin, Joseph Lister, Wilhelm Wundt, Louis Pasteur, and Robert Koch. Yet, until the end of the century, most of the significant research was done outside the walls of higher educational institutions. In Great Britain, for instance, it was the Royal Society and other such societies that fostered advanced studies and encouraged research. The basic curriculum of colleges and universities remained nontechnical and nonprofessional. The English cardinal John Henry Newman, lecturing in Dublin on *The Idea of a University* in 1852, stated that the task of the university was broadly to prepare young men "to fill any post with credit, and to master any subject with facility." The university ought not to attempt professional and technical education.

Condition of 18th- and 19th-century higher education

While Newman's words epitomized the views held in most of Europe and America, some of the new universities in Germany were moving toward the expansion of the educational enterprise. In 1807 Fichte had drawn up a plan for the new University of Berlin, which Humboldt two years later was able to realize in its founding. The school was dedicated to the scientific approach to knowledge, to the combination of research and teaching, and to the proliferation of academic pursuits; and its ideal was adopted in the founding or reconstitution of other universities—Breslau (1811), Bonn (1818), Munich (1826). By the third quarter of the 19th century the influence of German *Lernfreiheit* (freedom of the student to choose his own program) and *Lehrfreiheit* (freedom of the professor to develop the subject and to engage in research) was felt throughout the academic world. The unity of the universities, for better or worse, was more and more dissolved by the fragmentation of subjects into different branches. Some critics would eventually condemn what they considered to be the excesses of the free elective system and the extreme departmentalization of research and curricula. Much of the debate, however, would centre on the general education of undergraduates. In the meantime, the conviction, fathered in Germany, that research is a responsibility of universities was to inspire the founders of universities in the United States in the late 19th century.

Lernfreiheit and *Lehrfreiheit*

France. In France the Jesuit schools and the schools of other teaching orders created at the time of the Renaissance had reconciled the teaching of the new humanism with the established doctrines of the Roman Catholic church and flourished with special brilliance. But, despite the changes brought about by the Renaissance and the attention given to the sciences in the 17th and 18th centuries, it was not until the advent of the French Revolution that the universal right to education was proclaimed (1791).

That principle was compromised when Napoleon came to power, however. Although he maintained that the matter of education was an important issue and thought that a common culture with common ideals was essential to nation-building, he felt that, from a political standpoint, the bourgeoisie and upper classes were most important. His national education system therefore served children of those classes. This led to reorganization of the structure of

Napoleonic system of education

secondary and higher education in a unified state system, with secondary schools maintained by the communes, and with state lycées, universities, and special institutions of higher education. Within this structure the rector of a university headed a teaching body, recruited by the state and supervised by an inspectorate, ranging through various grades up to the university council. Grades of proficiency in studies, from simple certificates to the degrees of *baccalauréat, licence,* and doctorate were awarded on the result of examinations, and these tests were made a necessary condition of entry into such professions as medicine, law, and teaching. This structure, despite many modifications, has survived until modern times.

Development of state education. French educational history in the 19th century is essentially the story of the struggle for the freedom of education, of the introduction at the secondary level of the modern and scientific branches of learning, and, under the Third Republic, of the establishment of primary education, at once secular and compulsory, between the ages of six and 12. There were also a middle education between the ages of 13 and 16 and, finally, a professional and technical education.

Under the restoration of the monarchy in 1814, education fell inevitably under the control of the church; but, during the bourgeois monarchy of Louis Philippe, a law was passed in 1833 that laid the foundations of modern primary instruction, obliging the communes to maintain schools and pay the teachers. The higher primary schools that were founded were suppressed by Roman Catholic conservatives in 1850 (their restoration later constituted one of the great positive services rendered by the Third Republic to the cause of popular education). The 1850 law restored the "liberty of teaching" that, in effect, meant free scope for priestly schools, but it also made provision for separate communal schools for girls, for adult classes, and for the technical instruction of apprentices. In 1854 France

Centralization of French education

was divided for purposes of educational administration into 16 districts called *académies,* each administered by a rector and each with a university at the apex of the educational structure. The rector not only was made the chief administrator of the university but also was responsible for secondary and higher education within his *académie;* he nominated candidates for administrative positions in his area, appointed examination committees, supervised examination content and procedures, and presided over an academic council. Unlike the political division in some other countries, the *académies* were given little power or authority of their own; rather, they were administrative arms of the national ministry of education.

After the Franco-Prussian War, the Third Republic addressed itself to the organization of primary instruction as "compulsory, free, and secular." The law of 1878 imposed on communes the duty of providing school buildings and provided grants-in-aid. The national government also henceforth paid salaries throughout the public sector of education. In 1879 a law was passed compelling every department to maintain training colleges for male and female teachers. The law of 1881 abolished fees in all primary schools and training colleges; the law of 1882 established compulsory attendance; and, finally, the law of 1886 enacted that none but lay persons should teach in the public schools and abolished in those schools all distinctively religious teaching.

Secondary education. In European systems of education, secondary education was preeminently a preparation for the university, with aims and ideals of general culture that differentiated it radically and at the very outset from education of the elementary type. Down to the beginning of the 20th century, the French system could be regarded as a typical and extreme example of the European theory.

European dual plan

The characteristic European organization has been called the dual plan: elementary and secondary education were distinct types, and only a minority of the elementary-school pupils passed on to the secondary schools, generally only if they were bright and could win scholarships through a competitive examination. The secondary schools were of two kinds: lycées and communal colleges. The lycées, maintained by tuition fees and state scholarships, taught the ancient languages, rhetoric, logic, ethics, mathematics, and physical science. The communal colleges, established by municipalities or individuals and maintained by tuition fees, offered a partial lycée curriculum, featuring Latin, French, mathematics, history, and geography. Pupils who did not complete a secondary education program generally entered civil service or other white-collar occupations. With the development of commerce and industry in the 19th century, France instituted the *écoles primaires supérieures,* or "higher primary schools," for those who did not go on to universities but who needed a better education than the primary schools could give. The curricula of these schools were somewhat more advanced than those of the primary schools; pupils remained longer (up to the age of 16) and were prepared for employment in business as white-collar workers but generally at a lower level than pupils who came from the lycées. In effect, the different types of schools tended to maintain class cleavages since students of the secondary schools enjoyed higher social and occupational prestige than those of the upper primary schools.

The foundation of secondary schools for girls was in its way one of the most notable achievements of the Third Republic. It was inaugurated by the law of Dec. 22, 1880, called after its author the Loi Camille Sée. Until World War II, the curricula were different from those of the boys' schools, and the course of study was only five years. There were no ancient languages, and mathematics was not carried to so high a level as in the boys' lycées.

England. Influenced by doctrines of laissez-faire, England hesitated a long time before allowing the state to intervene in educational affairs. At the beginning of the 19th century, education was regarded as entirely the concern of voluntary or private enterprise, and there was much unsystematic philanthropy. Attempts were made to channel and concentrate it, and many hoped that the Church of England and the dissenting churches would join in a concerted effort to provide a national system of elementary education on a voluntary basis. But discordant views prevented such cooperation, and two voluntary societies were founded, one representative of the Church of England and the other of dissent. In 1829 the Roman Catholics were emancipated by law from disabilities they had long suffered, and so they also were able to provide voluntary schools. Other religious bodies joined in the effort to meet the growing need for elementary schools, but it was soon evident that voluntary finance would not be equal to this formidable task. In 1833 the government made a small building grant to these societies, and in this modest way state intervention began. Six years later a committee of the Privy Council was established to administer the state grants, now made annually, and to arrange for the inspection of voluntary schools aided from public funds. The work involved led to the establishment of a small central education department, which was the beginning of the ministry of education.

Laissez-faire and English voluntary education

Matthew Arnold was influential in pressing upon the English conscience the importance of public education for the state. While serving as inspector of elementary schools from 1851 to 1886, he studied European school systems and contrasted the meagre educational contributions of the English state with the more generous ones of Continental states.

Elementary Education Act. England prolonged its reliance on voluntary initiative for year after year as population increased, and, with the growing industrialization, people crowded into the new towns. At last in 1870 Parliament, after long, acrimonious debates, passed an Elementary Education Act, the foundation upon which the English educational system has been built. Religious teaching and worship were the crucial issues in the debates, and the essentials of the settlement agreed upon were (1) a dual system of voluntary and local-authority schools and (2) careful safeguards to ensure as far as possible that no child would receive religious teaching that was at variance with the wishes of his parents. It was left to the school boards—as these first local education authorities were called—to decide on an individual basis whether to make elementary education compulsory in their districts. In 1880, however, it was made compulsory throughout

England and Wales, and in 1891 fees were abolished in all but a few elementary schools.

Secondary and higher education. Secondary education, however, was still left to voluntary and private enterprise. Attention was focused on the "public" schools (independent secondary schools such as Eton and Harrow, usually for boarders from upper- and well-to-do middle-class homes), which under the leadership of outstanding headmasters such as Thomas Arnold were thoroughly reformed. As headmaster of Rugby School (1828–42), Arnold is credited with changing the face of public education in England by instilling a spirit of moral responsibility and intellectual integrity grounded in Christian ethics. Arnold's aims of school life—religious and moral principles, gentlemanly conduct, and intellectual ability—where to have an enduring influence on the English public-school system.

Several new universities were founded during the 19th century, and the latter half of it saw the founding of a number of girls' high schools and boarding schools offering an education that was comparable to that available in boys' public schools and grammar schools. Several training colleges for teachers were established by voluntary agencies, and universities and university colleges toward the end of the century undertook the training of postgraduates as teachers in departments of education created for this purpose.

Russia. Influenced by the disintegration of the serf system, the trend toward industrialization and modernization, and the democratic ideas of the French Revolution, Tsar Alexander I at the beginning of the 19th century tried to institute new educational reforms. The statutes of 1803 and 1804 followed the pattern set by Peter I the Great and Catherine II the Great in the 18th century for utilitarian, scientific, and secular education. The old Catherinian schools were remodeled and new schools founded. Schools were to be free and under state control. Rural peasants were to be taught reading, writing, arithmetic, and elements of agriculture at the parochial schools (*prikhodskiye uchilishcha*); pupils in the district schools of urban areas (*uyezdnye uchilishcha*) and the provincial schools (*gimnazii*) were to be prepared for careers as civil servants or for other white-collar occupations (law, political economy, technology, and commerce). The elementary and secondary schools were integrated with the universities.

Russian reform and reaction

Nicholas I, coming to the throne in 1825, considered this democratic trend harmful and decreed that:

> It is necessary that in every school the subjects of instruction and the very methods of teaching should be in accordance with the future destination of pupils, that nobody should aim to rise above that position in which it is his lot to remain.

A new statute of 1828 decreed that parochial schools were intended for the peasants, the district schools for merchants and other townspeople, and *gimnazii* for children of the gentry and civil servants. Instruction in the *gimnazii* in Latin and Greek was increased. Although the legislation of Nicholas I established a class system, the utilitarian character of the whole system remained.

The Russian radical intelligentsia was fiercely opposed to the privileged schools for the gentry and demanded the reestablishment of a democratic system with a more modern curriculum in secondary schools. This was coupled with the demand for the emancipation of the serfs and the equality of women in education. The new tsar in 1855, Alexander II, inaugurated a period of liberal reforms. The serfs were emancipated in 1861, and thus all social restrictions were removed. A new system of local government in rural areas (*zemstvo*) was enacted with a right to found schools for the peasantry, now free. Combined efforts of the government, *zemstva,* and peasant communities produced a growth of schools in the rural areas. The utilitarian trend was evident in the establishment of technical schools with vocational differentiation. The education of women was promoted, and the first higher courses for women were founded in main cities.

The reign of Alexander II, which was later marked by reactionary measures and political oppression, ended in his assassination in 1881 by the terrorist branch of the Narodniki revolutionary organization. A period of reaction followed under his successor, Alexander III. All reforms were suspended, and the growth of educational institutions was interrupted. The chief procurator of the Holy Synod attempted to build up a rival system of parochial schools under the control of the orthodox clergy; and the minister of public instruction tried to return to the class system of Nicholas I. These reactionary measures set back the growth of education. Four-fifths of all children were deprived of education. The result was that at the turn of the century nearly 70 percent of Russia's male population and 90 percent of its female population were illiterate (1897 census). The aboriginal dwellers of Russia's national outskirts (more than one-half of the country's population) were almost totally illiterate. (J.J.Ch.)

The United States. Administered locally everywhere, schooling of the United States's masses in the republic's younger days was immensely diverse. In New England, primary schooling enjoyed public support. In the South, apart from supplying a meagre learning to pauper children, the states abstained from educational responsibility. In the middle states elementary schools were sometimes public; more often they were parochial or philanthropic. Only beyond the Alleghenies was there any federal provision for education. There, under the Articles of Confederation, the Ordinance of 1787 reserved a plot of land in every prospective township for the support of education. The measure not only laid the groundwork for education in the states of the Ohio Valley and the Great Lakes, it also became a precedent for national educational aid. Thus, in 1862 the Morrill Act granted every state establishing a public agricultural college 30,000 acres (12,000 hectares) of public land for each of its lawmakers in Congress. Since then some 12 million acres (five million hectares) have been distributed, on which some 70 of the so-called land-grant colleges currently flourish.

American land-grant colleges

Several of the Founding Fathers expressed belief in the necessity of public education, but only Thomas Jefferson undertook to translate his conviction into actuality. Convinced that democracy can be effective only in the hands of an enlightened people, he offered Virginia's lawgivers a plan in 1779 to educate schoolchildren at public cost for three years and a few gifted boys beyond that. The proposal encountered resistance from both the ruling classes and the clergy; they regarded instruction as a private or an ecclesiastical prerogative. Jefferson's plan was rejected, as was another he submitted some 40 years later. Although his ideas enlightened educational thought throughout the country, only one of Jefferson's dreams reached actuality in his lifetime: the University of Virginia opened in 1825, the most up-to-date institution of its sort, the first frankly secular university in America and the closest to a modern-day conception of a state university.

The educational awakening. When Jefferson died in 1826 the nation stood on the threshold of a stupendous transformation. During the ensuing quarter century it expanded enormously in space and population. Old cities grew larger and new ones more numerous. The era saw the coming of the steamboat and the railroad. Commerce flourished and so did agriculture. The age witnessed the rise of the common man with the right to vote and hold office. It was a time of overflowing optimism, of dreams of perpetual progress, moral uplift, and social betterment.

The common school

Such was the climate that engendered the common school. Open freely to every child and upheld by public funds, it was to be a lay institution under the sovereignty of the state, the archetype of the present-day American public school. Bringing the common school into being was not easy. Against it bulked the doctrine that any education which excluded religious instruction—as all state-maintained schools were legally compelled to do—was godless. Nor had there been any great recession of the contention that education was not a proper governmental function and for a state to engage therein was an intrusion into parental privilege. Still more distasteful was the fact that public schooling would occasion a rise in taxes.

Yet the common school also mustered some formidable support, and finally, in 1837, liberal Massachusetts lawmakers successfully carried through a campaign for a

state board of education. It is especially to Horace Mann, the board's first secretary, that Massachusetts credits its educational regeneration. To gather data on educational conditions in Massachusetts, Mann roved the entire commonwealth. He lectured and wrote reports, depicting his dire findings with unsparing candour. There were outcries against him, but when Mann resigned, after 12 years, he could take pride in an extraordinary achievement. During his incumbency, school appropriations almost doubled. Teachers were awarded larger wages; in return they were to render better service. To help them Massachusetts established three state normal schools, the first in America. Supervision was made professional. The school year was extended. Public high schools were augmented. Finally, the common school, under the authority of the state, though still beset by difficulties, slowly became the rule.

What Mann accomplished in Massachusetts, Henry Barnard (1811–1900) achieved in Connecticut and Rhode Island. More reserved than Mann, Barnard has come down the ages as the "scholar of the educational awakening." He became the first president of the Association for the Advancement of Education and editor of its *American Journal of Education,* in whose 30 volumes he discussed virtually every important pedagogical idea of the 19th century.

Similar campaigns were under way in other areas. In Pennsylvania the assault centred on the pauper school; in New York it was against sectarianism. On the westward-moving frontier, old educational ideas and traditions had to compete in an environment antagonistic to privilege and permanence. There was controversy everywhere, however, over the state's right to assume educational authority and especially its power to levy school taxes. Future handling of this issue in the West was foretold in 1837, when Michigan realized a state-supported and state-administered system of education in which the state university, the University of Michigan under the leadership of Henry Tappan, played an integral part.

The rise of the American high school

Secondary education. Once the common school was solidly entrenched, the scant opportunity afforded the lower classes for more than a rudimentary education fell under increasing challenge. If it was right to order children to learn reading, writing, and arithmetic and to offer them free tax-supported schooling, some reasoned, then it was also right to accommodate those desiring advanced instruction. Before long, a few common schools, yielding to parental insistence, introduced courses beyond the elementary level. Such was the germ of the high school in the U.S.

The first high school in the United States opened in Boston in 1821 as the English Classical School, a designation that soon was changed to English High School. Designed for the sons of the "mercantile and mechanic classes," it provided three years of free instruction in English, mathematics, surveying, navigation, geography, history, logic, ethics, and civics. In 1825 New York City inaugurated the first high school outside New England. The next year Boston braved free secondary education for girls, judiciously diluted and restricted to 130. When the number of applicants vastly exceeded this figure, the city fathers abandoned the project.

The high-school movement was spurred less by these diffuse developments than by legislation by Massachusetts in 1827 that ordered towns of 500 families to furnish public instruction in American history, algebra, geometry, and bookkeeping, in addition to the common primary subjects. Furthermore, towns of 4,000 were to offer courses in history, logic, rhetoric, Latin, and Greek. The measure lacked public backing, but it set the guideposts for similar legislation elsewhere. The contention that government had no right to finance high schools remained an issue until the 1870s, when Michigan's supreme court, finding for the city of Kalamazoo in litigation brought by a taxpayer, declared the high school to be a necessary part of the state's system of public instruction.

Education for females. Though the common school vouchsafed instruction to girls, girls' chances to attend high school—not to say college—were slight. The "female academies," attended mainly by daughters of the middle class, were not numerous, and they varied in their emphases, often stressing social or domestic subjects. The truth is that as late as the 1840s, when the lowliest man could vote and hold office, women were haltered by taboos of every sort. But as America advanced industrially, and more and more women flocked to the mill and the office, their desire for greater educational opportunity grew. As in the struggle for the common school, the cause of women's education bred leaders, many of whom founded schools and communicated internationally. In 1833 Oberlin College in Ohio hazarded coeducation, and 20 years later Antioch College, also in Ohio, followed suit. Beyond the Mississippi every state university, except that of Missouri, was coeducational from its beginning. The East moved more warily; Cornell University was the first Eastern school to become coeducational, in 1872.

Co-education

Higher education. While women were crusading for greater educational opportunity, the college itself was undergoing alteration. It had begun as a cradle of divinity, but, as the 18th century waned, it was displaying a mounting secularity. In the course of the 19th century, not only did colleges surge in number, but some of the more enterprising of them undertook to reshape their purpose. Soon after its opening in 1885, Bryn Mawr College in Pennsylvania announced courses for the master's and doctor's degrees. Inspired by the scholarly accomplishments of German universities, Johns Hopkins University in Baltimore, founded in 1867, put its weight on research. Twenty years later Clark University in Massachusetts opened as a purely graduate school. Soon the graduate trend invaded older schools as well.

The early normal schools, or teacher-training schools, were primitive; often they were merely higher elementary schools, rehearsing their students for a year in basic reading and arithmetic, rectitude and piety, some history, mathematics, and physiology, and, if they survived, a rudimentary pedagogy. After the 1860s the ideas and experiments of Pestalozzi and Froebel combined with widespread social-democratic influences on education and advances in psychological thought to change schooling. This confluence, which was most noticeable in elementary education, resulted in the appearance of the kindergarten and in methods proceeding from the nature of the child and including content representing more of the present society. While much of the rationale was religious or mystical, the outcome was socially and psychologically more realistic. Since the early phases of schooling were initially the only concern of teacher training, it was natural that the idea of preparing teachers to use techniques derived from the new concepts, including the greater systematization introduced by Herbart, and the necessity for teachers to learn specifically about the child would substantially augment teacher-training programs and lay the groundwork for immense institutional expansion in the first half of the 20th century. (A.E.M./R.F.L.)

The British dominions. *Canada.* In the early period of the 19th century, until about 1840, schooling in Canada was much the same as it was in England; it was provided through the efforts of religious and philanthropic organizations and dominated by the Church of England. Although there was overlap among types of schools (identified historically), there are records of parish schools, charity schools, Sunday schools, and monitorial schools for the common people. The instructional fare was a rudimentary combination of religious instruction and literacy skills, perhaps supplemented by some practical work.

More advanced education was limited to the upper social classes and was given in Latin grammar schools or in private schools with various curricular extensions on the classical base. Academies, largely supported by the middle class of nonconformist groups, presented a broad curriculum of liberal arts that spanned the secondary and higher levels of education. In general, instruction relied on a simple chain concept of "transmission-absorption-mental storage," which was kept going by direct application of reward or punishment.

In the middle period, which lasted to about 1870, public systems of education emerged, accommodating religious interests in a state framework. Public support was won

for the common school, leading toward universal elementary education. Secondary and higher education began to assume a public character. The principle of local responsibility under central provincial authority was elaborated in the respective provinces.

Church and state relations in Canada

Of central importance in the development of Canadian education is the kind of agreement reached on church–state relations in education during this period. At one extreme is the arrangement made in Newfoundland from 1836 to accommodate all numerically represented denominations separately within a loose system (not until 1920 was a unified system of education developed, which still works through five denominational subsystems); at the other extreme are the arrangements made in British Columbia, which became decisive when it entered the Canadian Confederation, to establish and maintain a free, unified, centralized nonsectarian system. Other provinces eventually developed patterns that represented compromises. The Nova Scotia–New Brunswick pattern, for instance, provided a unified system that in principle was nonsectarian but that allowed the grouping of Roman Catholic children for education, thus legalizing sectarian schools within the system. Ontario placed separate Catholic schools within a unified school system. Québec supported a dual confessional system from the 1840s to the 1960s, with parallel structures for Roman Catholic and Protestant schooling at both the local and provincial levels. Manitoba adopted Québec's dual confessional system in 1871, then changed to a unified, centralized nonsectarian system amid much controversy in 1896.

The British North America Act of 1867, Canada's constitution, lodged authority for education in the provinces, at the same time guaranteeing denominational rights (in the "minority-school protective clause") if such rights existed by law at the time of entry into confederation. These two provisions established the pluralistic nature of Canadian education, and the union of the provinces and the entrance of western provinces gave Canada, by 1880, a national base on which to build the Canadian institution of education.

The final years of the 19th century were years of structural formalization of the educational foundations developed in the productive middle period. In this, Ontario's leadership was evident, especially as it affected the model of education evolving in the western territories. After Alberta and Saskatchewan were admitted as provinces in 1905, some divergence from Ontario took place: notably, both provinces required that Roman Catholic taxes go to separate Catholic schools (the decision in Ontario was based on free choice), and Alberta allowed separate school privileges through the secondary level. (Saskatchewan extended full funding of Roman Catholic separate schools to the end of high school in the early 1960s, Ontario in the late 1980s).

Toward the end of the 19th century, elementary schooling, by then established, was becoming compulsory. The cost of secondary education was diminishing, and the distinction in level and curriculum between the secondary and the elementary school was sharpened in the system of public schools. Communities were responsible for maintaining schools through a combination of local taxes and provincial grants, while provincial departments standardized the conduct of schooling through inspections, examinations, and prescription of course content and materials.

Changes in instructional theory, taking place during the latter part of the 19th century throughout the Western world, revolutionized the classroom. One major shift was from the imposition of knowledge on the mind of the learner to an emphasis on the learner's activity of perception and comprehension of knowledge. The impact of science on the higher-school curriculum was matched by its impact on educational theory and, consequently, on teacher training. Both scientific disciplines (such as educational psychology) and scientific methods of teaching became necessary to the training of teachers who were to operate in a new setting of teacher–pupil and subject-matter relations.

Australia. The development of Australian education through the 19th century was affected by a pervasive British influence, by a continuous economic struggle against harsh environmental conditions, and by the tendency for population to be concentrated in centres that accrued and extended political authority over the region. The particular historical thread around which educational developments took place was the question of denominational schools.

Church and state relations in Australia

From the first immigrant landing in 1788 through the early decades of the 19th century, education was provided on an occasional and rather haphazard basis, by the most expedient means available. In general, the assumption and the practice was that schooling would be provided by the church or by church organizations, such as the SPGFP, and colonial governments made small grants to aid such provision. It was also assumed that the Church of England would dominate the religious-educational scene, and a Church and School Corporation was set up in 1826 to administer endowments for Church of England efforts. Even at this early stage, however, the resistance of Nonconformists, especially Presbyterians and Roman Catholics, shortly defeated the attempt to "establish" Church of England institutions. The only early organized attempt at mass education was through monitorial systems.

In 1833 the governor of New South Wales asserted government responsibility for education by proposing the introduction of a nondenominational system that would reduce religion in schools to reading commonly approved scriptures and to providing release time for sectarian instruction by clergymen. The importance of the proposal lay in its spirit of religious compromise and its initiation of state responsibility for education, both of which were predictive of future development.

Because of sectarian resistance, mainly from Anglican and Catholic groups, so-called national schools were introduced alongside denominational schools in 1848 as a dual system, administered by two corresponding boards. Through the middle period of the century, similar sectarian compromises were found in other Australian colonies. The establishment of state systems were, however, seriously impeded by the extremity of the struggle for survival in hostile geographic conditions. In New South Wales a Public Schools Bill was passed in 1866, creating a single Council of Education. State aid to denominational schools was continued but under conditions stipulated by the state.

Victoria became a separate colony in 1850 and was initially fraught with particular problems occasioned by the arrival of a migrant gold-rush population. Little was accomplished in education, other than increased assistance to religious denominations, until 1856. After that the move for a state system gained impetus, and a Common Schools Bill was passed in 1862, establishing a system similar to that accepted in New South Wales. Soon after separation, Queensland's Primary Education Bill was passed in 1860, subordinating denominational schools and reinforcing the principle of common-school development in Australia. South Australia held to a continuous development of a general system based on common Christianity, but Western Australia's Elementary Education Bill of 1871 returned to dual support for both government and voluntary schools.

The support for state educational systems increased during the 1860s and 1870s as an alternative to interdenominational conflict was sought. In this development the Protestants, gradually and sometimes reluctantly, acquiesced. Catholic resistance was never overcome, and the consequent evolution of a separate Roman Catholic school system did not diminish Catholic dissatisfaction with the movement to state schools. The dilemma of Catholic citizens with regard to nonsectarian public education was universal: as citizens they were financially obligated for the public schools; as Roman Catholics they were committed to education in schools of their own faith.

The intention to educate all children and to raise the quality of instruction in common schools required governmental actions that could transform voluntary, exclusive, uneven provisions into uniform public standards. In Australia, particular motivating factors were the dramatic increases in population and economic growth and the recognized inadequacy of existing schools. The establishment

of secular public-school systems under government control was made unequivocal through the passage of legislation between 1872 and 1895. These bills did not abolish general Christian instruction, nor did they generally refuse release time for sectarian instruction. They did disallow sectarian claims for financial support and for a place in public education. The decision was for the operation of schools for all children, undertaken by the one agency that could act on behalf of the whole society, the government.

New Zealand. In New Zealand's early colonial period, between 1840 and 1852, certain provisions were made for endowments for religious and educational purposes, but education was considered, in accordance with prevailing views in England, a private or voluntary matter. Corresponding to general social distinctions, academic education was relegated to denominational, fee-charging schools, and common education was provided as a charitable service. Religious preference was avoided as much as possible, with the aim of minimizing sectarian conflict.

Church and state relations in New Zealand

Secular opposition to religious bias, even on a pluralistic basis, was, however, already evident. In 1852 New Zealand was granted self-government under the Constitution Act, and responsibility for education was placed in the councils of the six provinces. Although each province acted independently and somewhat according to the traditions of the dominant cultural group, the general sentiment moved in the next 20 years toward the establishment of public school systems. By 1876, when the provincial governments were abolished, the people of New Zealand, through varying regional decisions, had accepted governmental responsibility for education, had opted for nonsectarian schools, and had started on the path to free, compulsory common schooling.

The basic national legislation was passed in 1877. The Education Act provided for public elementary education that would be secular, free to age 15, and compulsory to age 13. Because of enforcement difficulties and legal exceptions, the compulsory clause was rather loose, but it instituted the rule. It was strengthened between 1885 and 1898, and high-school enrollments increased steadily after 1911. The act of 1877 also revised the administrative structure under a national ministerial Department of Education. Initially, the central department was little more than a funding source, while critical control was vested in regional boards elected by local school committees. In the competitive struggle between the department and the regional boards that waxed and waned well into the 20th century, neither gained the exclusive dominance sometimes sought. The primary position of the central authority in educational administration was confirmed in the reform period between 1899 and 1914, however, when control of inspectors, effective control of primary teacher appointment and promotion, and stipulative control in fund granting went to the Department of Education. These developments, together with curriculum and examination reforms, marked a new beginning in New Zealand education. (R.F.L.)

THE SPREAD OF WESTERN EDUCATIONAL PRACTICES TO ASIAN COUNTRIES

India. Originally the British went to India as tradesmen, but gradually they became the rulers of the country. On Dec. 31, 1600, the East India Company was established, and, like all commercial bodies, its main objective was trade. Gradually during the 18th century the pendulum swung from commerce to administration; the deterioration of Mughal power in India, the final expulsion of French rivals in the Seven Years' War, and the virtual appropriation of Bengal and Bihār in a treaty of 1765 had all made the company a ruling power. In spite of this, the company did not recognize the promotion of education among the natives of India as a part of its duty or obligation. For a long time the British at home were greatly opposed to any system of public instruction for the Indians, as they were for their own people.

Education under the East India Company

The feelings of the public authorities in England were first tested in the year 1793, when William Wilberforce, the famous British philanthropist, proposed to add two clauses to the company's charter act of that year for sending out schoolmasters to India. This encountered the greatest opposition in the council of directors, and it was found necessary to withdraw the clauses. For 20 years thereafter, the ruling authorities in England refused to accept responsibility for the education of Indian people. It was only in 1813, when the company's charter was renewed, that a clause was inserted requiring the governor-general to devote not less than 100,000 rupees annually to the education of Indians.

Some organization was required in order to disburse the educational grant. A General Committee of Public Instruction, constituted in Calcutta in 1823, started its work with an Orientalist policy, rather than a Western-oriented one, since the majority of the members were Orientalists. The money available was spent mainly on the teaching of Sanskrit and Arabic and on the translation of English works into these languages. Some encouragement was also given to the production of books in English.

Meanwhile, a new impetus was given to education from two sources of different character. One was from the Christian missionaries and the other from a "semirationalist" movement. The Christian missionaries had started their educational activities as early as 1542, upon the arrival of St. Francis Xavier. Afterward the movement spread throughout the land and exercised a lasting influence on Indian education. It gave a new direction to elementary education through the introduction of instruction at regular and fixed hours, a broad curriculum, and a clear-cut class system. By printing books in different vernaculars, the missionaries stimulated the development of Indian languages. But hand in hand with the study of the vernaculars went the teaching of Western subjects through the medium of English, called in India "English education."

Besides the missionaries, there were men in Bengal who, though admitting the value of Oriental learning for the advancement of civilization, thought that better things could be achieved through the so-called English education. In 1817 these semirationalists, led by Rām Mohan Roy, the celebrated Indian reformer, founded the Hindu College in Calcutta, the alumni of which established a large number of English schools all over Bengal. The demand for English education in Bengal thus preceded by 20 years any government action in that direction.

In the meantime the influence of the Orientalists was waning in the General Committee, as younger members with more radical views joined it. They challenged the policy of patronizing Oriental learning and advocated the need for spreading Western knowledge through the medium of English. Thus arose the controversy as to whether educational grants should be used to promote Oriental learning or Western knowledge. The controversy between the Orientalists and the Anglicists was decided in favour of the latter by the famous Minute on Education of 1835 submitted by Thomas Babington Macaulay, the legal member of the governor-general's executive council. His recommendations were accepted by Lord William Bentinck, the governor-general. The decision was announced on March 7, 1835, in a brief resolution that determined the character of higher education in India for the ensuing century. Although the schools for Oriental learning were maintained for some years, the translation of English books into Sanskrit and Arabic was immediately discontinued. Thus the system of "English education" was adopted by the government. It should be noted, however, that primary education did not attract any attention at all.

Orientalists versus Anglicists

Bentinck's resolution was followed by other enactments accelerating the growth of English education in India. The first was the Freedom of Press Act (1835), which encouraged the printing and publication of books and made English books available at low cost. Two years later, Persian was abolished as the language of record and the courts (to the dismay of the Muslims) and was replaced by English and Indian languages in higher and lower courts, respectively. Finally, Lord Hardinge, as governor-general, issued a resolution on Oct. 10, 1844, declaring that for all government appointments preference would be given to the knowledge of English. These measures strengthened the position of English in India, and the lingering prejudices against learning English vanished forever.

Although English education held its ground in Bengal, the Bengal government did not neglect vernacular education altogether. Moreover, in Bombay, Madras, and the North-Western Provinces there was as yet little effective demand for English, and the tendency was to lay the main stress on Indian languages. Bombay adopted the policy of encouraging primary education and spreading Western science and knowledge through the mother tongue. This was done under the able guidance of Mountstuart Elphinstone, then the governor, even though the government also conducted an English school in almost every district in the province. Between 1845 and 1848 a bitter controversy arose regarding the language of instruction, but the issue was between the mother tongue and English, and not between a classical language and English as it was in Bengal. The controversy gathered strength every day; and, in those days of centralization, the matter had to be referred to the Bengal government, which advised the Bombay government to concentrate its attention on English education alone, thus throttling the growth of education through the mother tongue in Bombay. Meanwhile, the Madras government was biding its time, leaving the field of positive effort open to Christian missionaries; as a result of this missionary initiative, English education in the Madras presidency was more extensively imparted than in Bombay.

A laudable experiment in the field of vernacular education was carried out by Lieutenant Governor James Thomason in the North-Western Provinces. Thomason's *ḥalqabandī* system attempted to bring primary education within easy reach of the common people. In each *ḥalqah* (circuit) of villages, a school was established in the most central village so that all the villagers within a radius of two miles might avail themselves of it. For the maintenance of these schools the village landholders agreed to contribute at the rate of 1 percent of their land income. The experiment proved successful, and in 10 years Thomason opened 897 schools and provided elementary education for 23,688 children.

Ḥalqabandī system

The next step in the history of Indian education is marked by Sir Charles Wood's epoch-making Dispatch of 1854, which led to (1) the creation of a separate department for the administration of education in each province, (2) the founding of the universities of Calcutta, Bombay, and Madras in 1857, and (3) the introduction of a system of grants-in-aid. Even when the administration of India passed from the East India Company into the hands of the British crown in 1858, Britain's secretary of state for India confirmed the educational policy of Wood's Dispatch.

The newly established universities did not initially undertake any teaching responsibilities but were merely examining bodies. Their expenses were confined to administration and could be met from the fees paid by the candidates for their degrees and certificates. Although the establishment of the universities did result in a rapid expansion of college education and although the products of the new learning displayed keen scholarship, the value of learning nevertheless soon decayed. In such circumstances it was ironic for the Indian Education Commission of 1882 to declare, "The university degree has become an accepted object of ambition, a passport to distinction in public services and in the learned professions." Another undesirable practice was the domination of the universities over secondary education through their entrance examinations. University policies regarding curricula, examination systems, language of instruction, and other vital problems began to be chalked out by university teachers who had little experience in schoolteaching and who kept the administrative needs and requirements of colleges in the forefront. Thus, secondary schools increasingly prepared their students for a college education and not for life in general.

Excessive emphasis on higher education

The new system also became top-heavy. It must be stated that the commission of 1882 made a very valuable recommendation that the "elementary education of the masses, its provision, extension and improvement requires strenuous efforts of the state in a still larger measure than heretofore." It also desired to check the wild race for academic distinction and "to divert some part of the rapidly swelling stream of students into channels of a more practical character." Despite this warning, however, alternative courses in commerce, agriculture, and technical subjects that were offered in a limited number of selected schools did not prove popular. The educated classes could not be diverted from their conventional path.

In a general view of education during the last two decades of the 19th century, drift was more apparent than government resolve. Elementary education was starved and undernourished, and secondary education was suffering from want of proper supervision. There was an unplanned growth of high schools and colleges since the Education Commission had given a free charter to private enterprise. Many of these private institutions were "coaching institutions rather than places of learning." The universities had no control over them, and state control was negligible because the government had adopted a laissez-faire policy.

The second half of the 19th century is, nonetheless, of great significance to the country because modern India may indeed be said to be a creation of this period. It brought about a renaissance by breaking down geographic barriers and bringing different regions and long-separated Indian communities into close contact with one another. The blind admiration for Western culture was gradually passing away, and a new vision and reorientation in thought were coming about. A feeling of dissatisfaction also developed toward the existing governmental and missionary institutions. It was felt by some of the Indian patriots that the character of Indian youths could be built by Indians themselves. This led to the establishment of a few notable institutions aiming at imparting sound education to Indian youth on national lines—institutions such as the Anglo-Mohammedan Oriental College in Alīgarh (1875), the D.A.V. College in Lahore (1886), and the Central Hindu College in Vārānasi (1898). The politically minded classes of the country had also come to regard education as a national need. They were critical of the government's educational policy and resented any innovation that might restrain the pace of educational advance or diminish liberty. (S.N.M.)

Japan. *The Meiji Restoration and the assimilation of Western civilization.* In 1867 the Tokugawa (Edo) shogunate, a dynasty of military rulers established in 1603, was overthrown and the imperial authority of the Meiji dynasty was restored, leading to drastic reforms of the social system. This process has been called the Meiji Restoration, and it ushered in the establishment of a politically unified and modernized state.

In the following generation Japan quickly adopted useful aspects of Western industry and culture to enhance rapid modernization. But Japan's audacious modernization would have been impossible without the enduring peace and cultural achievements of the Tokugawa era. It had boasted a high level of Oriental civilization, especially centring on Confucianism, Shintōism, and Buddhism. The ruling samurai had studied literature and Confucianism at their *hankō* (domain schools); the commoners had learned reading, writing, and arithmetic at numerous *terakoya* (temple schools). Both samurai and commoners also pursued medicine, military science, and practical arts at *shijuku* (private schools). Some of these schools had developed a fairly high level of instruction in Western science and technology by the time of the Meiji Restoration. This cultural heritage helped equip Japan with a formidable potential for rapid Westernization. Indeed, some elements of Western civilization had been gradually introduced into Japan even during the Tokugawa era. The shogunate, notwithstanding its isolationist policy, permitted trade with the Dutch, who conveyed modern Western sciences and arts to Japan. After 1853, moreover, Japan opened its door equally to other Western countries, a result of pressures exerted by the United States Navy under Admiral Matthew C. Perry. Thenceforth, even before the Meiji Restoration, Japanese interest in foreign languages became intense and diverse.

The Tokugawa heritage

Western studies, especially English-language studies, became increasingly popular after the Restoration, and Western culture flooded into Japan. The Meiji government dispatched study commissions and students to Europe and to the United States, and the so-called Westernizers

defeated the conservatives who tried in vain to maintain allegiance to traditional learning.

Establishment of a national system of education. In 1871 Japan's first Ministry of Education was established to develop a national system of education. Ōki Takatō, the secretary of education, foresaw the necessity of establishing schools throughout the nation to develop national wealth, strength, and order, and he outlined a strategy for acquiring the best features of Western education. He assigned commissioners, many of whom were students of Western learning, to design the school system, and in 1872 the Gakusei, or Education System Order, was promulgated. It was the first comprehensive national plan to offer schooling nationwide, according to which the nation was divided into eight university districts, which were further divided into 32 middle-school districts, each accommodating 210 primary-school districts. Unlike the class-based schooling offered during the Tokugawa period, the Gakusei envisioned a unified, egalitarian system of modern national education, designed on a ladder plan. Although the district system was said to have been borrowed from France, the new Japanese education was based on the study of Western education in general and incorporated elements of educational practice in all advanced countries. Curricula and methods of education, for instance, were drawn primarily from the United States.

Gakusei, or Education System Order

This ambitious modern plan for a national education system fell short of full realization, however, because of the lack of sufficient financial support, facilities and equipment, proper teaching materials, and able teachers. Nevertheless, the plan represented an unprecedented historic stage in Japanese educational development. Under the Gakusei system, the Ministry of Education, together with local officials, managed with difficulty to set up elementary schools for children aged six to 14. In 1875 the 24,000 elementary schools had 45,000 teachers and 1,928,000 pupils. This was achieved by gradually reorganizing *terakoya* in many areas into modern schools. The enrollment rate reached only 35 percent of all eligible children, however, and no university was erected at all.

In 1873 David Murray, a professor from the United States, was invited to Japan as an adviser to the Ministry of Education; another, Marion M. Scott, assumed direction of teacher training and introduced American methods and curricula at the first normal school in Tokyo, established under the direct control of the ministry. Graduates of the normal school played an important role in disseminating teacher training to other parts of the country. By 1874 the government had set up six normal schools, including one for women. The normal school designed curricula for the primary schools, modeled after those of the United States, and introduced textbooks and methods that spread gradually into the elementary schools of many regions.

The work of educators and teachers from abroad

The conservative reaction. Following the repression of the Satsuma Rebellion, a samurai uprising in 1877, Japan again forged ahead toward political unity, but there was an increasing trend of antigovernment protest from below, which was epitomized by the Movement for People's Rights. Because of the Satsuma Rebellion, the government was in heavy financial difficulties. Also, with the people's inclination toward Western ideas fading away, a conservative reaction began to emerge, calling for a revival of the Confucian and Shintō legacies and a return to local control of education as practiced in the pre-Restoration era.

Discontent had been mounting among the rural people against the Education System Order of 1872, mainly because it had imposed upon them the financial burdens of establishing schools and yet had not lived up to expectations. Another cause of dissatisfaction was a sense of irrelevance that Japanese attributed to schooling largely based on Western models. The curriculum developed according to the 1872 order was perceived to have little relation to the social and cultural needs of that day, and ordinary Japanese continued to favour the traditional schooling of the *terakoya.* Tanaka Fujimaro, then deputy secretary of education, just returning from an inspection tour in the United States, insisted that the government transfer its authority over education to the local governments, as in the United States, to reflect local needs in schooling. Thus, in 1879 the government nullified the Gakusei and put into force the Kyōikurei, or Education Order, which made for rather less centralization. Not only did the new law abolish the district system that had divided the country into districts, it also reduced central control over school administration, including the power to establish schools and regulate attendance. The Kyōikurei was intended to encourage local initiatives. Such a drastic reform to decentralize education, however, led to an immediate deterioration of schooling and a decline in attendance in some localities; criticism arose among those prefectural governors who had been striving to enforce the Gakusei in their regions.

As a countermeasure, the government introduced a new education order in 1880 calling for a centralization of authority by increasing the powers of the secretary of education and the prefectural governor. Thereafter, the prefecture would provide regulations within the limits of criteria set by the Ministry of Education; some measure of educational unity was thus reached on the prefectural level, and the school system received some needed adjustment. Yet, because of economic stagnation, school attendance remained low.

Conservatism in education gained crucial support when the Kyōgaku Seishi, or the Imperial Will on the Great Principles of Education, was drafted by Motoda Nagazane, a lecturer attached to the Imperial House in 1870. It stressed the strengthening of traditional morality and virtue to provide a firm base for the emperor. Thereafter, the government began to base its educational policy on the Kyōgaku Seishi with emphasis on Confucian and Shintōist values. In the elementary schools, *shūshin* (national moral education) was made the all-important core of the curricula, and the ministry compiled a textbook with overtones of Confucian morality.

Reemphasis on Confucian and Shintōist values

Establishment of nationalistic education systems. With the installation of the cabinet system in 1885, the government made further efforts to pave the way for a modern state. The promulgation of the Meiji constitution, the constitution of the empire of Japan, in 1889 established a balance of imperial power and parliamentary forms. The new minister of education, Mori Arinori, acted as a central figure in enforcing a nationalistic educational policy and worked out a vast revision of the school system. This set a foundation for the nationalistic educational system that developed during the following period in Japan. Japanese education thereafter, in the Prussian manner, tended to be autocratic.

Based on policies advocated by Mori, a series of new acts and orders were promulgated one after another. The first was the Imperial University Order of 1886, which rendered the university a servant of the state for the training of high officials and elites in various fields. Later that year orders concerning the elementary school, the middle school, and the normal school were issued, forming the structural core of the pre-World War II education system. The ministry carried out sweeping revisions of the normal-school system, establishing it as a completely independent track, quite distinct from other educational training. It was marked by a rigid, regimented curriculum designed to foster "a good and obedient, faithful, and respectful character." As a result of these reforms the rate of attendance at the four-year compulsory education level reached 81 percent by 1900.

Together with these reforms, the Imperial Rescript on Education (Kyōiku Chokugo) of 1890 played a major role in providing a structure for national morality. By reemphasizing the traditional Confucian and Shintō values and redefining the courses in *shūshin,* it was to place morality and education on a foundation of imperial authority. It would provide the guiding principle for Japan's education until the end of World War II.

Promotion of industrial education. Ever since the Meiji Restoration in 1868, the national target had been *fukoku-kyōhei* ("wealth accumulation and military strength") and industrialization. From the outset the Meiji government had been busy introducing science and technology from Europe and America but nevertheless had difficulties in realizing such goals.

Inoue Kowashi, who became minister of education in 1893, was convinced that modern industries would be the most vital element in the future development of Japan and thus gave priority to industrial and vocational education. In 1894 the Subsidy Act for Technical Education was published, followed by the Technical Teachers' Training Regulations and the Apprentice School Regulations. The system of industrial education was in general consolidated and integrated. These measures contributed to the training of many of the human resources required for the subsequent development of modern industry in Japan.

(A.Na./N.S.)

Education in the 20th century

SOCIAL AND HISTORICAL BACKGROUND

International wars, together with an intensification of internal stresses and conflicts among social, racial, and ideological groups, have characterized the 20th century and have had profound effects on education. Rapidly spreading prosperity but widening gaps between rich and poor, immense increases in world population but a declining birth rate in Western countries, the growth of large-scale industry and its dependence on science and technological advancement, the increasing power of both organized labour and international business, and the enormous influence of both technical and sociopsychological advances in communication, especially as utilized in mass media, are changes that have had far-reaching effects. Challenges to accepted values, including those supported by religion; changes in social relations, especially toward versions of group and individual equality; and an explosion of knowledge affecting paradigms as well as particular information mark a century of social and political swings, always toward a more dynamic and less categorical resolution. The institutional means of handling this uncertain world have been to accept more diversity while maintaining basic forms and to rely on management efficiency to ensure practical outcomes.

The two world wars weakened the military and political might of the larger European powers. Their replacement by "superpowers" whose influence did not depend directly on territorial acquisition and whose ideologies were essentially equalitarian helped to liquidate colonialism. As new independent nations emerged in Africa and Asia and the needs and powers of a "third world" caused a shift in international thinking, education was seen to be both an instrument of national development and a means of crossing national and cultural barriers. One consequence of this has been a great increase in the quantity of education provided. Attempts have been made to eradicate illiteracy, and colleges and schools have been built everywhere.

The growing affluence of masses of the population in high-income areas in North America and Europe has brought about, particularly since World War II, a tremendous demand for secondary and higher education. Most children stay at school until 16, 17, or even 18 years of age, and a substantial fraction spend at least two years at college. The number of universities in many countries doubled or trebled between 1950 and 1970, and the elaboration of the tertiary level continues.

Demands of industry for technical and continuing education

This growth is sustained partly by the industrial requirements of modern scientific technology. New methods, processes, and machines are continually introduced. Old skills become irrelevant; new industries spring up. In addition, the amount of scientific, as distinct from merely technical, knowledge grows continually. More and more researchers, skilled workers, and high-level professionals are called for. The processing of information has undergone revolutionary change. The educational response has mainly been to develop technical colleges, to promote adult education at all levels, to turn attention to part-time and evening courses, and to provide more training and education within the industrial enterprises themselves.

The adoption of modern methods of food production has diminished the need for agricultural workers, who have headed for the cities. Urbanization, however, brings problems: city centres decay, and there is a trend toward violence. The poorest remain in these centres, and it becomes difficult to provide adequate education. The radical change to large numbers of disrupted families, where the norm is a single working parent, affects the urban poor extensively but in all cases raises an expectation of additional school services. Differences in family background, together with the cultural mix partly occasioned by change of immigration patterns, requires teaching behaviour and content appropriate to a more heterogeneous school population.

MAJOR INTELLECTUAL MOVEMENTS

Influence of psychology and other fields on education. The attempt to apply scientific method to the study of education dates back to the German philosopher Johann Friedrich Herbart, who called for the application of psychology to the art of teaching. But not until the end of the 19th century, when the German psychologist Wilhelm Max Wundt established the first psychological laboratory at the University of Leipzig in 1879, were serious efforts made to separate psychology from philosophy. Wundt's monumental *Principles of Physiological Psychology* (1874) had significant effects on education in the 20th century.

Influence of William James

William James, often considered the father of American psychology of education, began about 1874 to lay the groundwork for his psychophysiological laboratory, which was founded officially at Harvard in 1891. In 1878 he established the first course in psychology in the United States and in 1890 published his famous *The Principles of Psychology,* in which he argued that the purpose of education is to organize the child's powers of conduct so as to fit him to his social and physical environment. Interests must be awakened and broadened as the natural starting points of instruction. James's *Principles* and *Talks to Teachers on Psychology* cast aside the older notions of psychology in favour of an essentially behaviourist outlook; they asked the teacher to help educate heroic individuals who would project daring visions of the future and work courageously to realize them.

James's student Edward L. Thorndike is credited with the introduction of modern educational psychology, with the publication of *Educational Psychology* in 1903. Thorndike attempted to apply the methods of exact science to the practice of psychology. James and Thorndike, together with the American philosopher John Dewey, helped to clear away many of the fantastic notions once held about the successive steps involved in the development of mental functions from birth to maturity.

Interest in the work of Sigmund Freud and the psychoanalytic image of the child in the 1920s, as well as attempts to apply psychology to national training and education tasks in the 1940s and '50s, stimulated the development of educational psychology, and the field has become recognized as a major source for educational theory. Eminent researchers in the field have advanced knowledge of behaviour modification, child development, and motivation. They have studied learning theories ranging from classical and instrumental conditioning and technical models to social theories and open humanistic varieties. Besides the specific applications of measurement, counseling, and clinical psychology, psychology has contributed to education through studies of cognition, information processing, the technology of instruction, and learning styles. After much controversy about nature versus nurture and about qualitative versus quantitative methods, Jungian, phenomenological, and ethnographic methods have taken their place alongside psychobiological explanations to help educationists understand the place of heredity, general environment, and school in development and learning.

The relationship between educational theory and other fields of study has become increasingly close. Social science may be used to study interactions and speech to discover what is actually happening in a classroom. Philosophy of science has led educational theorists to attempt to understand paradigmatic shifts in knowledge. The critical literature of the 1960s and '70s attacked all institutions as conveyors of the motives and economic interests of the dominant class. Both social philosophy and critical sociology have continued to elaborate the themes of social control and oppression as embedded in educational institutions. In a world of social as well as intellectual

change, there are necessarily new ethical questions, such as those dealing with abortion, biological experimentation, and child rights, which place new demands on education and require new methods of teaching.

Traditional movements. Against the various "progressive" lines of 20th-century education, there have been strong voices advocating older traditions. These voices were particularly strong in the 1930s, in the 1950s, and again in the 1980s. Essentialists stress those human experiences that they believe are indispensable to people living today or at any time. They favour the "mental disciplines" and, in the matter of method and content, put effort above interest, subjects above activities, collective experience above that of the individual, logical organization above the psychological, and the teacher's initiative above that of the learner.

Essentialist, liberal, and religious education

Closely related to essentialism is what used to be called humanistic, or liberal, education in its traditional form. Although many intellectuals have argued the case, Robert M. Hutchins, president and then chancellor of the University of Chicago from 1929 to 1951, and Mortimer J. Adler, professor of the philosophy of law at the same institution, are its most recognized proponents. Adler argued for the restoration of an Aristotelian viewpoint in education. Maintaining that there are unchanging verities, he sought a return to education fixed in content and aim. Hutchins denounced American higher education for its vocationalism and "anti-intellectualism," as well as for its delight in minute and isolated specialization. He and his colleagues urged a return to the cultivation of the intellect.

Opposed to the fundamental tenets of pragmatism is the philosophy that underlies all Roman Catholic education. Theocentric in its viewpoint, Catholic scholasticism has God as its unchanging basis of action. It insists that without such a basis there can be no real aim to any type of living, and hence there can be no real purpose in any system of education. The church's

> whole educational aim is to restore the sons of Adam to their high position as children of God. [It insists that] education must prepare man for what he should do here below in order to attain the sublime end for which he was created. (From Pius XI, encyclical on the "Christian Education of Youth," Dec. 31, 1929.)

Everything in education—content, method, discipline—must lead in the direction of man's supernatural destiny.

New foundations. The three concerns that guided the development of 20th-century education were: the child, science, and society. The foundations for this trilogy were laid by so-called progressive education movements supporting child-centred education, scientific-realist education, and social reconstruction.

Progressive education. The progressive education movement was part and parcel of a broader social and political reform called the Progressive movement, which dates to the last decades of the 19th century and the early decades of the 20th. Elementary education had spread throughout the Western world, largely doing away with illiteracy and raising the level of social understanding. Yet, despite this progress, the schools had failed to keep pace with the tremendous social changes that had been going on.

Dissatisfaction with existing schools led several educational reformers who wished to put their ideas into practice to establish experimental schools during the last decade of the 19th century and in the early 20th century. The principal experimental schools in America until 1914 were the University of Chicago Laboratory School, founded in 1896 and directed by John Dewey; the Francis W. Parker School, founded in 1901 in Chicago; the School of Organic Education at Fairhope, Ala., founded by Marietta Johnson in 1907; and the experimental elementary school at the University of Missouri (Columbia), founded in 1904 by Junius L. Meriam. The common goal of all was to eliminate the school's traditional stiffness and to break down hard and fast subject-matter lines. Each school adopted an activity program. Each operated on the assumption that education was something that should not be imposed from without but should draw forth the latent possibilities from within the child. And each believed in the democratic concept of individual worth.

Some early experimental schools

Dewey, whose writings and lectures influenced educators throughout the world, laid the foundations of a new philosophy that continues to affect the whole structure of education, particularly at the elementary level. His theories were expounded in *School and Society* (1899), *The Child and the Curriculum* (1902), and *Democracy and Education* (1916). For Dewey, philosophy and education render service to each other. Education becomes the laboratory of philosophy. Society should be interpreted to the child through daily living in the classroom, which acts as a miniature society. Education leads to no final end; it is something continuous, "a reconstruction of accumulated experience," which must be directed toward social efficiency. Education is life, not merely a preparation for life.

The views of John Dewey

The influence of progressive education advanced slowly during the first decades of the 20th century. Nevertheless, a number of progressive schools were established, including the Play School and the Walden School in New York City, the Shady Hill School in Cambridge, Mass., the Elementary School of the University of Iowa, and the Oak Lane Day School in Philadelphia. Helen Parkhurst's Dalton Plan, introduced in 1920 at Dalton, Mass., pioneered individually paced learning of broad topics. Carleton Washburne's Winnetka Plan, instituted in 1919 at Winnetka, Ill., viewed learning as a continuous process guided by the child's own goals and capabilities. The Gary Plan, developed in 1908 at Gary, Ind., by William Wirt, established a "complete school," embracing work, study, and play for all grades on a full-year basis.

The spread of progressive education became more rapid from the 1920s on and was not confined to any particular country. In the United States the Progressive Education Association (PEA) was formed in 1919. The PEA did much to further the cause of progressive education until it ended, as an organization, in 1955. In 1921 Europe's leading progressives formed the New Education Fellowship, later renamed the World Education Fellowship.

The notions expressed by progressive education have influenced public-school systems everywhere. Some of the movement's lasting effects can be seen in the activity programs, imaginative writing and reading classes, projects linked to the community, flexible classroom space, dramatics and informal activities, discovery methods of learning, self-assessment systems, and programs for the development of citizenship and responsibility found in school systems all over the world.

Child-centred education. Proponents of the child-centred approach to education have typically argued that the school should be fitted to the needs of the child and not the child to the school. These ideas, first explored in Europe, notably in Rousseau's *Émile* (1762) and in Pestalozzi's *How Gertrude Teaches Her Children* (1801), were implemented in American systems by pioneering educators such as Francis W. Parker. Parker became superintendent of schools in Quincy, Mass., in 1875. He assailed the mechanical, assembly-line methods of traditional schools and stressed "quality teaching," by which he meant such things as activity, creative self-expression, excursions, understanding the individual, and the development of personality.

A different approach to child-centred education arose as a result of the study and care of the physically and mentally handicapped. Teachers had to invent their own methods to meet the needs of such children, because the ordinary schools did not supply them. When these methods proved successful with handicapped children, the question arose whether they might not yield even better results with ordinary children. During the first decade of the 20th century, the educationists Maria Montessori of Rome and Ovide Decroly of Brussels both successfully applied their educational inventions in schools for ordinary boys and girls.

The Montessori method's underlying assumption is the child's need to escape from the domination of parent and teacher. According to Montessori, children, who are the unhappy victims of adult suppression, have been compelled to adopt defensive measures foreign to their real nature in the struggle to hold their own. The first move toward the reform of education, therefore, should be directed toward educators: to enlighten their consciences, to

The Montessori method

remove their perceptions of superiority, and to make them humble and passive in their attitudes toward the young. The next move should be to provide a new environment in which the child has a chance to live a life of his own. In the Montessori method, the senses are separately trained by means of apparatuses calculated to enlist spontaneous interest at the successive stages of mental growth. By similar self-educative devices, the child is led to individual mastery of the basic skills of everyday life and then to schoolwork in arithmetic and grammar.

The Decroly method can be characterized as a program of work based on centres of interest and educative games. Its basic feature is the workshop-classroom, in which children can go freely about their own occupations. Behind the complex of individual activities there is a carefully organized scheme of work based on an analysis of the fundamental needs of the child. The principle of giving priority to wholes rather than to parts is emphasized in teaching children to read, write, and count, and care is taken to reach a comprehensive view of the experiences of life.

The Montessori and the Decroly methods have spread throughout the world and have widely influenced attitudes and practices of educating young children.

Pestalozzian principles have also encouraged the introduction of music education into early childhood programs. Research has shown that music has an undeniable effect on the development of the young child, especially in such areas as movement, temper, and speech and listening patterns. The four most common methods of early childhood music education are those developed by Émile Jaques-Dalcroze, Carl Orff, and Zoltán Kodály and the Comprehensive Musicianship approach. The Dalcroze method emphasizes movement; Orff, dramatization; Kodály, singing games; and Comprehensive Musicianship, exploration and discovery. Another popular method, developed by the Japanese violinist Shinichi Suzuki, is based on the theory that young children learn music in the same way that they learn their first language.

Scientific-realist education. The scientific-realist education movement began in 1900 when Édouard Claparède, then a doctor at the Psychological Laboratory of the University of Geneva, responded to an appeal from the women in charge of special schools for backward and abnormal children in Geneva. The experience brought him to realize some of the defects of ordinary schools. Not as much thought is given, he argued, to the minds of children as is to their feet. Their shoes are of different sizes and shapes, made to fit their feet. When shall we have schools to measure? The psychological principles needed to adapt education to individual children were expounded in his *Psychologie de l'enfant et pédagogie experimentale* (1909). Later Claparède took a leading part in the creation of the J.-J. Rousseau Institute in Geneva, a school of educational sciences to which came students from all over the world.

Jean Piaget's studies of child development

Theorists such as Claparède hoped to provide a scientific basis for education, an aim that was furthered by the Swiss psychologist Jean Piaget, who studied in a philosophical and psychological manner the intellectual development of children. Piaget argued, on the basis of his observations, that development of intelligence exhibits four chief stages and that the sequence is everywhere the same, although the ages in the stages of development may vary from culture to culture.

The first stage takes place during infancy, when children, even before they learn to speak, put objects together (addition), then separate them (subtraction), perceiving them as collections, rings, networks, groups. By the age of two or three, a basis has been laid. The children have developed kinetic muscular intelligence to some degree—they can think with their fingers, their hands, their bodies. Aided by language, the capacity for symbolic thinking slowly develops. This constitutes the second stage. Up to the age of seven or eight, some of the fundamental categories of adult thinking are still absent: there is seldom any notion, for instance, of cause and effect relationships.

The third stage is that of concrete operation. The child has begun to know how to deal with mental symbols and acquires abstract notions such as "responsibility." But the child operates only when in the presence of concrete objects that can be manipulated. Pure abstract thinking is still too difficult. Teaching at this stage must be exceedingly concrete and active; purely verbal teaching is out of place. Only after about 12 years of age, with the onset of adolescence, do children develop the power to deal with formal mental operations not immediately attached to objects. Only then do theories begin to acquire real significance, and only then can purely verbal teaching be used.

The child's total development, particularly emotional and social growth, also concerned educational reformers. They pointed out the error in assuming that incentives to mental effort are the same for adults and children. The English philosopher Alfred North Whitehead, in his doctrine of the "Cycle of Interests," put forward a theory in line with the ideas of the reformers. Romance, precision, and generalization, said Whitehead, are the stages through which, rhythmically, mental growth proceeds.

> Education should consist in a continual repetition of such cycles. Each lesson in a minor way should form an eddy cycle issuing in its own subordinate process.

Whitehead believed that any scheme of education must be judged by the extent to which it stimulates a child to think. From the beginning of education, children should experience the joy of discovery.

Social-reconstructionist education. Social-reconstructionist education is based on the theory that society can be reconstructed through the complete control of education. The objective is to change society to conform to the basic ideals of the political party or government in power or to create a utopian society through education.

Communist education is probably the most pervasive version of operational social-reconstructionism in the world today. Originally based on the philosophy of Karl Marx and institutionalized in the Soviet Union, it now reaches a large proportion of the world's youth. From the 1950s onward, much attention has been paid to the ideal of "polytechnization." Man, so the argument runs, is not simply *Homo sapiens* but rather *Homo faber,* the constructor and builder. He attains full mental, moral, and spiritual development through entering into social relations with others, particularly in cooperative efforts to produce material, artistic, and spiritual goods and achievements. The school should prepare pupils for such productive activities—for instance, by studying and, if possible, sharing in the work done in field, farm, or factory.

Reconstructionist education in Israel

A different social-reconstructionist movement is that of the kibbutzim (collective farms) of Israel. The most striking feature of kibbutz education is that the parents forgo rearing and educating their offspring themselves and instead hand the children over to professional educators, sometimes immediately after birth. The kibbutzim type of education developed for both practical and economic reasons, but gradually educational considerations gained prominence. These were: (1) that the kibbutz way of life makes for complete equality of the sexes, (2) that the education of children in special children's houses is the best way of perpetuating the kibbutz way of life, (3) that collective education is more "scientific" than education within the family, inasmuch as children are reared and trained by experts (*i.e.,* qualified nurses, kindergarten teachers, and other educators), in an atmosphere free of the tensions engendered by family relationships, and (4) that collective education is more democratic than traditional education and more in keeping with the spirit of cooperative living.

MAJOR TRENDS AND PROBLEMS

The idea of social-reconstructionist education rests on a 19th-century belief in the power of education to change society. In the last quarter of the 20th century there has been considerable pessimism, but the idea that schooling can influence either society or the individual is widely held, affecting the growth of tertiary-level alternatives, management strategies, and education of disadvantaged people, both in industrialized and in developing countries.

Integrative trends

The international concern with assistance to people in the non-Western world has been paralleled by the inclusiveness that has characterized education in the 20th century. Education has been seen as a primary instrument in recognizing and providing equality for those suffering

disadvantage because of sex, race, ethnic origin, age, or physical disability. This has required revisions of textbooks, new consciousness about language, and change in criteria for admission to higher levels. It has led to more demanding definitions of equality involving, for example, equality of outcome rather than of opportunity.

The inclusion of all children and youth is part of a general integrative trend that has accelerated since World War II. It relates to some newer developments as well. Concern for the earth's endangered environment has become central, emphasizing in both intellectual and social life the need for cooperation rather than competition, the importance of understanding interrelationships of the ecosystem, and the idea that ecology can be used as an organizing concept. In a different vein, the rapid development of microelectronics, particularly the use of computers for multiple functions in education, goes far beyond possibilities of earlier technological advances. Although technology is thought of by some as antagonistic to humanistic concerns, others argue that it makes communication and comprehension available to a wider population and encourages "system thinking," both ultimately integrative effects.

The polarization of opinion on technology's effects and most other important issues is a problem in educational policy determination. In addition to the difficulties of governing increasingly large and diverse education systems, as well as those of meeting the never-ending demands of expanding education, the chronic lack of consensus makes the system unable to respond satisfactorily to public criticism and unable to plan for substantive long-range development. The political and administrative responses so far have been (1) to attend to short-run efficiency by improving management techniques and (2) to adopt polar responses to accommodate polar criticisms. Thus, community and community schools have been emphasized along with central control and standardization, and institutional alternatives have been opened, while the structure of main institutions has become more articulated. For example, the focus of attention has been placed on the transition stages, which earlier were virtually ignored: from home to school, from primary to secondary to upper secondary, from school to work. Tertiary institutions have been reconceived as part of a unified level; testing has become more sophisticated and credentials have become more differentiated either by certificate or by transcript. Alternative teaching strategies have been encouraged in theory, but basic curriculum uniformity has effectively restricted the practice of new methods. General education is still mainly abstract, and subject matter, though internally more dynamic, still rests on language, mathematics, and science. There has been an increasing reliance on the construction of subject matter to guide the method of teaching. Teachers are entrusted with a greater variety of tasks, but they are less trusted with knowledge, leading political authorities to call for upgrading of teacher training, teacher in-service training, and regular assessment of teacher performance.

Recent reform efforts have been focused on integrating general and vocational education and on encouraging lifelong or recurrent education to meet changing individual and social needs. Thus, not only has the number of students and institutions increased, as a result of inclusion policies, but the scope of education has also expanded. This tremendous growth, however, has raised new questions about the proper functions of the school and the effectiveness for life, work, or intellectual advancement of present programs and means of instruction.

WESTERN PATTERNS OF EDUCATION

The United Kingdom. *Early 19th to early 20th century.* English education has been less consciously nationalist than that of continental European countries, but it has been deeply influenced by social class structure. Traditionally, the English have held that the activity of the government should be restricted to essential matters such as the defense of property and should not interfere in education, which was the concern of family and church. The growth of a national education system throughout the 19th century continued without a clear plan or a national decision. The cornerstone of the modern system was laid by the Elementary Education Act of 1870, which accepted the principle that the establishment of a system of elementary schools should be the responsibility of the state. It did not, however, eliminate the traditional prominence of voluntary agencies in the sphere of English education. Nor did it provide for secondary education, which was conducted largely by voluntary fee-charging grammar schools and "public" schools. These public schools were usually boarding schools charging rather high fees. Their tradition was aristocratic, exclusive, formal, and classical. Their main goal was to develop "leaders" for service in public life. In 1900 one child in 70 could expect to enter a secondary school of some kind. The grammar schools copied the curriculum of the public schools, so that only the intellectual and social elite were able to attend.

The Elementary Education Act of 1870

In 1899 an advance was made toward the development of a national system encompassing both elementary and secondary education by creating a Board of Education as the central authority for education. The Balfour Act of 1902 established a comprehensive system of local government for both secondary and elementary education. It created new local education authorities and empowered them to provide secondary schools and develop technical education. The Education Act of 1918 (The Fisher Act) aimed at the establishment of a "national system of public education available for all persons capable of profiting thereby." Local authorities were called upon to prepare plans for the orderly and progressive development of education. The school-leaving age was raised to 14, and power was given to local authorities to extend it to 15.

Education Act of 1944. The Education Act of 1944 involved a thorough recasting of the educational system. The Board of Education was replaced by a minister who was to direct and control the local education authorities, thereby assuring a more even standard of educational opportunity throughout England and Wales. Every local education authority was required to submit for the minister's approval a development plan for primary and secondary education and a plan for further education in its area. Two central advisory councils were constituted, one for England, another for Wales. These had the power, in addition to dealing with problems set by the minister, to tender advice on their own initiative. The total number of education authorities in England and Wales was reduced from 315 to 146.

The educational systems of Scotland and Northern Ireland are separate and distinct from that of England and Wales, although there are close links between them. The essential features of the Education Act of 1944 of England and Wales were reproduced in the Education Act of 1945 in Scotland and in the Education Act of 1947 in Northern Ireland. There were such adaptations in each country as were required by local traditions and environment.

The complexity of the education system in the United Kingdom arises in part from the pioneer work done in the past by voluntary bodies and a desire to retain the voluntary element in the state system. The act of 1944 continued the religious compromise expressed in the acts of 1870 and 1902 but elaborated and modified it after much consultation with the parties concerned. The act required that, in every state-aided primary and secondary school, the day should begin with collective worship on the part of all pupils and that religious instruction should be given in every such school. As in earlier legislation there was, however, a conscience clause and another to ensure that no teacher should suffer because of religious convictions. Religious instruction continues to be given in both fully maintained and state-aided voluntary schools, and opportunities exist for religious training beyond the daily worship and minimum required instruction. In many schools the religious offering has become nondenominational, and in areas of high non-Christian immigrant population consideration may be given to alternative religious provision.

Structural changes under the 1944 act

Two fundamental reforms in the act of 1944 were the requirement of secondary education for all, a requirement that meant that no school fees could be charged in any school maintained by public authority; and replacement of the former distinction between elementary and higher ed-

ucation by a new classification of "three progressive stages to be known as primary education, secondary education, and further education." To provide an adequate secondary education in accordance with "age, ability, and aptitude," as interpreted by the Ministry of Education, three separate schools were necessary: the grammar school, modeled on elite public schools, the less intellectually rigorous secondary modern school, and the technical school. If, in exceptional circumstances, such provisions were made in a single school, then the school would have to be large enough to comprise the three separate curricula under one roof. Children were directed to the appropriate school at the age of 11 by means of selection tests.

The tripartite system of grammar, secondary modern, and technical schools did not, in fact, flourish. The ministry had never been specific about the proportion of "technically minded" children in the population, but, in terms of school places provided in practice, it was about 5 percent. Since, on the average, grammar-school places were available to 20 percent, this left 75 percent of the child population to be directed to the secondary modern schools for which the ministry advocated courses not designed to lead to any form of qualification.

The comprehensive movement. Selection procedures at the age of 11 proved to be the Achilles' heel of the grammar school–secondary modern system. Various developments contributed to the downfall of selection at 11: first, the examination successes of the secondary modern schoolchildren; second, the failure of a significant proportion of the children so carefully selected for grammar schools; third, the report of a committee appointed by the British Psychological Society supported arguments that education itself promotes intellectual development and that "intelligence" tests do not in fact measure genetic endowment but rather educational achievement.

The main issue in the 1950s and '60s was whether or not the grammar schools should be retained with selection at 11 plus. One of the main arguments used was that the right of "parental choice" must be upheld. Another was that it was in the "English tradition" to retain a selective system. But gradually the number of comprehensive (nonselective) schools increased.

Policies of Labour and Conservative governments

The Labour Party during the election of 1964 promised to promote the establishment of the comprehensive school and to abolish selection at 11 plus. On taking office, however, the Labour government, instead of legislating, issued a circular in the belief that this would enlist local support and encourage local initiative. The result was conflict between national policy and local policy in some areas. The Conservative government elected in 1970 declared its intention of leaving decisions about reorganization to the local authorities. The comprehensive principle has since become dominant, and the number of comprehensive schools has grown under both Labour and Conservative governments so that most state-maintained secondary schools are now comprehensive. The administrative compromise of leaving organizational options open to local authorities has permitted variations to continue, however. Five to 6 percent of the school population attend completely independent private schools. Enrollment at the exclusively academic, often prestigious, and costly independent secondary schools may be preceded by attendance at private preparatory schools.

The primary school begins at age five and is usually divided into an infant stage (ages five to seven) and a junior stage (ages eight to 11). In those few localities using a middle-school organization, children attend the middle school from age eight or nine to age 13 or 14. Preschool provision is uneven, but a great deal of innovation has taken place in ideas and practices of early-childhood learning. In the infant school children work together with their teacher. Children may be placed together vertically in the same class, like a family group. Play is considered an activity of central significance in the infant school. It is a vehicle for the child's motivation and learning, carefully structured to promote cognitive development. The teacher's job is to set the environment through organization of space, time, and materials; to encourage, guide, and stimulate; and to see that all children learn and develop independence and responsibility. Studies are interrelated, and the curriculum is flexible.

British educational administration

The compromise regarding school organization is representative of the British educational administration's attempt to balance local and national interests delicately. Local education authorities are responsible for basic school operations, and much of the professional responsibility is passed on to the school. This representation of community and professional interest is underscored in policy documents, such as the 1980 Education Act's stipulation that governing boards include at least two parent and two teacher representatives. Local education authorities maintain a professional administrative staff and administer school finances, which are funded primarily by government grants and local property taxes.

Ultimate authority for education is at the national level, with the Department of Education and Science (formerly the Ministry of Education) headed by the secretary of state for education and science. The department is the agent of governmental policy. It reaches schools through circulars and directives as well as through Her Majesty's Inspectors of Schools. The inspectors increasingly advise and report on the general condition of schooling.

Under the Conservative government of Margaret Thatcher emphasis has been placed on management efficiency. While decentralization has applied to operational decisions, the government has increasingly pushed for standardization of curriculum and streamlining of assessment procedures. Traditionally, curriculum had been decentralized to the extreme in the United Kingdom, being a matter of teacher's professional judgment, unified only informally (though effectively) through the influence of teacher training, publicized curriculum projects, textbook choices, and public examination syllabi. This resulted in a great deal of curriculum agreement in the common schooling period, narrowing to a secondary core to age 16, including a wide range of options in the comprehensive school, and different basic curricula in selective systems. Independent schools showed some variations, particularly in the requirement of Latin, and the upper secondary stage was characterized by specialization. Through the 1970s and '80s, however, there was central pressure on curriculum improvement in science, practical elements, technical and vocational education, and the relationship of education to economic life. Influential publications have proposed standardization of the curriculum nationally.

Probably the issue that has received the most attention has been the relationship of education to the economy, to industry, to work. Much of the impact of this attention has been on the post-compulsory sector. Schemes developed outside of the educational establishment are providing training for young school-leavers. The Technical and Vocational Education Initiative calls for local education authority cooperation with the Manpower Services Commission in the introduction of technical courses which span school and post-school training. Recent reforms to the examination and certification system exemplify the government's thrust toward improvement of the education–economy link, toward rationalization of the system, and toward coordinated, standardized assessment procedures.

Further education. Further education is officially described as the "post-secondary stage of education, comprising all vocational and nonvocational provision made for young people who have left school, or for adults." Further education thus embraces the vast range of university, technical, commercial, and art education and the wide field of adult education. It is this sector of education, which is concerned with education beyond the normal school-leaving ages of 16 or 18, that has experienced the most astonishing growth in the number of students.

In the 19th century the dominance of Oxford and Cambridge was challenged by the rise of the civic universities, such as London, Manchester, and Birmingham. Following the lead of the 18th-century German universities and responding to a public demand for increased opportunity for higher education, Britain's new civic universities quickly acquired recognition—not only in technological fields but also in the fine and liberal arts.

Many new post-school technical colleges were founded

in the early 20th century. The Fisher Act of 1918 empowered the local authorities to levy a rate (tax) to finance such colleges. The universities, on the other hand, received funds from the central government through the University Grants Committee, established in 1911 and reorganized in 1920, after World War I.

A new type of technical college was established in the 1960s—the polytechnic, which provides mainly technological courses of university level as well as courses of a general kind in the arts and sciences. Polytechnics are chartered to award degrees validated by a Council for National Academic Awards.

Thus, the tertiary level in the United Kingdom is made up of colleges of further education, technical colleges, polytechnics, and universities. The colleges offer full-time and part-time courses beyond compulsory-school level. Polytechnics and universities are mainly responsible for degrees and research. The innovative Open University, with its flexible admission policy and study arrangements, opened in 1971. It uses various media to provide highly accessible and flexible higher education for working adults and other part-time students. It serves as an organizational model and provides course materials for similar institutions in other countries.

The Open University

Changes in British education in the second half of the 20th century have, without changing the basic values in the system, extended education by population, level, and content. New areas for expansion include immigrant cultural groups and multicultural content, the accommodation of special needs, and the development of tools and content in the expanding fields of microelectronics.

Germany. *Imperial Germany.* The formation of the German Empire in 1871 saw the beginning of centralized political control in the country and a corresponding emphasis on state purposes for education. Although liberal and socialist ideas were discussed, and even practiced in experimental schools, the main features of the era were the continued systematization of education, which had progressed in Prussia from 1763, and the class-based division of schools. Education for the great bulk of the population stressed not only literacy but also piety and morality, vocational and economic efficiency, and above all obedience and discipline. The minority of citizens in the upper social and economic strata were educated in separate schools according to a classical humanist rationale of intelligence and fitness that equipped them to fill the higher positions in the Reich. Reform proposals in the last decade of the 19th century led to an overhaul of the education system, but the changes did not remove class privileges.

The *Volksschule* was universal, free, and compulsory. The fundamental subjects were taught along with gymnastics and religion, which held important places in the curriculum. Girls and boys were taught in separate schools except when it was uneconomical to do so. Boys usually received training in manual work, and girls in domestic science. Graduates of the *Volksschule* found it almost impossible to enter the secondary school, which was attended almost exclusively by graduates of private preparatory schools charging fees. The *Volksschule* led its students directly to work and was thus separate and parallel to the secondary-school program rather than sequential.

The *Volksschule*

Boys who, at the age of nine, were about to enter secondary school had to decide on one of the three types of schools, each offering a different curriculum. The traditional classical *Gymnasium* stressed Latin and Greek. The *Realgymnasium* offered a curriculum that was a compromise between the humanities and modern subjects. The *Oberrealschule* stressed modern languages and sciences. Although Kaiser William II threw his influence on the side of the modernists in 1890, the *Gymnasium* continued to overshadow the other two schools until after World War II.

Secondary schools for girls were recognized by Prussia in 1872 and were extended and improved in 1894 and again in 1908. These schools were fee-paying and were thus available chiefly to the upper social and economic strata. The course of instruction lasted 10 years, from six to 16. This 10-year school was called the *Lyzeum,* the first three years being preparatory. Beyond it was the *Oberlyzeum,* which was divided into two courses: the *Frauenschule,* which offered a two-year general course, and the *Lehrerinnenseminar,* which offered a four-year course for prospective elementary-school teachers. Girls who wanted a secondary-school education similar to that of the boys transferred at the age of 13 to the *Studienanstalt.*

Continuation schools for the working class augmented apprenticeship training with part-time education. They were the forerunners of the part-time vocational *Berufsschulen,* which continue today. Greatly influenced by the ideas of Georg Kerschensteiner, these schools increased in importance in the early 20th century. Between 1919 and 1938 they filled out the secondary sector to ensure attendance at some kind of school for all youth to the age of 18.

Weimar Republic. In no sphere of public activity did the establishment of the Weimar Republic after 1919 cause more creative discussion and more far-reaching changes than in that of education. A four-year *Grundschule* was established, free and compulsory for all children. It was the basic building block for all subsequent social liberalization in education. Besides the elementary subjects and religion, the child was instructed in drawing, singing, physical training, and manual work. The *Oberstufe,* the four upper classes of the elementary school, combined with the *Grundschule,* formed a complete whole. Most elementary schools thus provided an eight-year course of study. Intermediate schools (*Mittelschulen*) were established for children who wished a longer and more advanced elementary-school course and were able to pay modest fees.

The Weimar constitution preserved the religious tradition, which had been an essential part of the school curriculum in Germany since the Reformation. No pupil, however, could be compelled to study religion, and no teacher could be forced to teach it. Communities were accorded the right to establish schools in accordance with the particular religious beliefs of the pupils.

As regards secondary education, the Weimar Republic kept the prewar division of *Gymnasium, Realgymnasium,* and *Oberrealschule.* (There were three comparable schools for girls.) In addition, there was established the *Aufbauschule,* which was a six-year school following completion of the seventh year of the elementary school, and the *Deutsche Oberschule,* a nine-year school that required two modern foreign languages and stressed German culture.

Secondary schools in the Weimar period

Nazi Germany. After Adolf Hitler's accession to power in 1933, the Nazis set out to reconstruct Germany society. To do that, the totalitarian government attempted to exert complete control over the populace. Every institution was infused with National Socialist ideology and infiltrated by Nazi personnel in chief positions. Schools were no exception. Even before coming to power, Hitler in *Mein Kampf* had hinted at his plans for broad educational exploitation. The Ministry of Public Enlightenment and Propaganda exercised control over virtually every form of expression—radio, theatre, cinema, the fine arts, the press, churches, and schools. The control of the schools began in March 1933 with the issuing of the first educational decree, which held that "German culture must be treated thoroughly."

The Nazi government attempted to control the minds of the young and thus, among other means, intruded Nazi beliefs into the school curriculum. A major part of biology became "race science," and health education and physical training did not escape the racial stress. Geography became geopolitics, the study of the fatherland being fundamental. Physical training was made compulsory for all, as was youth labour service. Much of the fundamental curriculum was not disturbed, however.

Changes after World War II. Immediately after World War II the occupying powers (Britain, France, and the United States in the western zones and the Soviet Union in the east) instituted education programs designed to clean out Nazi influence and to reflect their respective educational values. These efforts were soon absorbed into independent German educational reconstruction. The Basic Law of the Federal Republic of Germany (West Germany) of May 1949 granted autonomy in educational matters to the *Länd* (state) governments. Although efforts to strengthen the federal government's presence have waxed and waned, *Länd* governments remain independent and divided along political lines on educational reforms.

The two main political issues dividing the states have always been confessional schooling and the tripartite division of secondary schooling, with conservative states like Bavaria and Baden-Württemberg on the one side and socially progressive states like Hessen and West Berlin on the other. After a 20-year period of reform discussion on these issues, marked by influential state or national proposals, the balance shifted in the mid-1970s to the conservatives, albeit with a great deal of internal liberalization. That is, confessional schools and confessional instruction in schools remained, but the latter was increasingly in ecumenical or ethical versions. This change, like others, has been supported by the presence of a large number of non-German children representing various cultural beliefs and behaviours. On the issue of dividing secondary schools, in spite of continued strong intellectual and political support from some quarters, the movement toward comprehensive schools has at least for the time being died out. Even where comprehensive schools (*Gesamtschulen*) exist, they usually incorporate separate secondary paths. Nevertheless, the effective extension of common schooling through an "orientation stage" between elementary and secondary schooling, the attempt to develop each level so that it better serves more youth, even if differentially, and the functional integration of school branches through curriculum reform and transfer possibilities all point to a comprehensiveness within the system.

The German school structure

Education is compulsory from age six to 18. In general, pupils spend four years in the elementary school (*Grundschule*), six years in one of the lower secondary branches, and two years in one of the upper secondary branches. The first two years of the lower secondary school constitute the "orientation stage." Long governed by entrance examination, the choice of secondary school is now made by the parents, although performance at the orientation stage, especially in the subjects of German, mathematics, and foreign language (English), influences decisions.

About 25 percent of secondary-school-age children enter the *Gymnasium,* which, with different academic emphases, remains the successor to its classical ancestor. Roughly 40 percent attend the nonselective *Hauptschule* ("main school"), which offers basic subjects to grade nine or 10 and is followed by apprenticeship with part-time vocational school or by full-time vocational school. Approximately 25 percent attend the *Realschule* (formerly *Mittelschule*), which offers academic and prevocational options. It leads to vocational school or technical school, which in turn lead to commercial, technical, or administrative occupations. The vocational-technical sector has always been given careful government and industry attention, and the network now includes a wide range of methods and content alternatives, with levels up to a university equivalent. All of these institutions encompass general education, theory of the trade or industrial field, and work practice. The schools can be reentered from work and can provide an alternative path to the university.

One of the means of coordinating differences among *Länd* systems has been through the Conference of the Cultural Ministers of the states, and one of the important resolutions of this body, in 1973, was for reform of the upper secondary stage. Attention has been given to equalizing opportunities at this stage. This has affected the *Gymnasium* by shifting much of the traditional load to the upper level. Although the first stage is still academically demanding, the foreign-language requirement is much more flexible, and many students now leave for work at the end of the 10th school year. The upper level is required to reach the *Abitur,* qualifying the student for university entrance. Although the range of subjects has been extended, courses have been diversified, and final achievement is now indicated by a cumulative point system. The upper level of the *Gymnasium* is characterized by breadth of knowledge at a high intellectual standard, including cultural essentials as well as an academic concentration, and thus still captures the German educational ideal.

Whether due to periodic change, German tradition, or inadequate understanding of the reform process, the educational system has irresistibly returned to basic principles. The incorporation of new alternatives and individual opportunities yields an open rather than a fundamentally changed system. This may be the best way for education to meet the major political themes of modern Germany: individual rights as the criterion of policy determination and the European community as the broader context of national development.

France. *The Third Republic.* The establishment of the Third Republic (1870) brought about the complete renovation of the French schools, in the process of which education became a national enterprise. In 1882 primary education was made compulsory for all children between the ages of six and 13. In 1886, members of the clergy were forbidden to teach in the public schools, and in 1904 the teaching congregations were suppressed. France had thus established a national free, compulsory, and secularized system of elementary schooling. (Although secularization was a necessary government strategy, it was also necessary to permit private Catholic schools, and these have continued to enroll a significant number of French children.)

In spite of the attempt to unify education through national purpose and centralized means, two parallel systems existed, that of the public elementary schools and higher primary schools and that of the selective, overwhelmingly intellectual secondary lycées and their preparatory schools. The lycées emphasized classical studies through the study of Greek and Latin. It was not until 1902 that this exclusive emphasis was challenged by a reform promoting the study of modern languages and science and not until the period between World Wars I and II that education was seen to have a vocational function, other than grossly in a social-class sense, and thus to require democratization.

Centralized administration in French education

The administration of education in France has remained highly centralized and has continued to be concerned with every aspect of national education, including curricula, syllabi, textbooks, and teacher performance. At the head of the system is the minister of national education, who is advised and assisted by a hierarchy of officials. The country is divided into 27 educational administrative areas, each known as an "academy." The chief education officer is the rector, the minister's most important representative, who administers the laws and regulations. The inspectorate, represented by regional inspectors under an *inspecteur d'académie* and by national inspectors, has extensive bureaucratic and supervisory powers.

Changes after World War II. Since 1946 education has been included in the plans developed by the central planning commission in France. In general, government has been friendly to educational development and reform. Student protests in the late 1960s caused an antagonistic reaction, however, and teacher resistance appears to work against many government reform initiatives. Government reform trends in recent years have been toward increasing administrative efficiency and accountability, meeting national economic needs through improved technological education, improving the articulation of system parts, opening the school to the community, and correcting inequalities, through both curricular and organizational provisions. Attention has been given not only to "socializing" the system but also to correcting inequalities suffered by French ethnic minorities and immigrant children, to amending social-geographic inequalities, and to increasing options for the handicapped, in both special schools and, after the mid-1970s, regular schools.

The French school structure

In 1947 a commission established to examine the educational system recommended a thorough overhauling of the entire school system. Education was to be compulsory from the age of six to 18. Schooling was to be divided into three successive stages: (1) six to 11, aimed at mastery of the basic skills and knowledge, (2) 11 to 15, a period of guidance to discover aptitudes, and (3) 15 to 18, a stage during which education was to be diversified and specialized. The system has since consistently developed from one featuring a common elementary school to one incorporating a progression into separate paths. Reforms have aimed to provide equality of educational experience at each stage and to create curricular conditions that further career advancement without abridging general education or forcing students to choose a profession prematurely.

Preschool education is given in the *école maternelle,* in

which attendance is voluntary from the age of two to six years. Education is compulsory between six and 16 years of age and is free. The five-year elementary school is followed by a four-year lower secondary school, the *collège unique,* which has been the object of much attention. The first two years at the *collège unique* constitute the observation cycle, during which teachers observe student performance; during the remaining two years, the orientation cycle, teachers offer guidance and assist pupils in identifying their abilities and determining a career direction.

At the upper secondary level, from age 15 to 18, students enter either the general and technological high school (*lycée d'enseignement général et technologique*), successor to the traditional academic high school, or the vocational senior high school (*lycée d'enseignement professionel*), encompassing a range of vocational-technical studies and qualifications. Students entering the former choose one of three basic streams the first year, then concentrate the next two years on one of five sections of study: literary-philosophical studies, economics and social science, mathematics and physical science, earth science and biological science, or scientific and industrial technology. The number of sections and particularly the number of technological options is scheduled to be expanded. There is a common core of subjects plus options in grades 10 and 11, but all subjects are oriented to the pupil's major area of study. In grade 12 the subjects are optional. The *baccalauréat* examination taken at the end of these studies qualifies students for university entrance. It consists of written and oral examinations. More than half of the 70 percent who pass are females. The proportion of the age group reaching this peak of school success has risen continuously, with corresponding effects on entrance to higher education.

Vocational-technical secondary education includes a number of options. Each of the courses leading to one of the 30 or so technical *baccalauréats* requires three years of study and gives access to corresponding studies in higher education. Students may also choose to obtain, in descending order of qualification requirements and course demands, the technician diploma (*brevet de technicien*), the diploma of vocational studies (*brevet d'études professionelles*), or the certificate of vocational aptitude (*certificat d'aptitude professionelle*). A one-year course conferring no particular qualification is also available, or youths may opt for apprenticeship training in the workplace.

French university education

Higher education is offered in universities, in institutes attached to a university, and in the *grandes écoles.* Students attend for two to five years and sit either for a diploma or, in certain establishments, for university degrees or for a competitive examination such as the agrégation. Undergraduate courses last for three or four years, depending on the type of degree sought.

The universities went through a period of violent student dissatisfaction in the late 1960s. Reforms ensued encouraging decentralization, diversification of courses, and moderation of the importance of examinations. Nevertheless, the failure or dropout rate in the first two years is still high, and there are marked differences in status among institutions and faculties.

Teachers are graded according to the results of a competitive academic examination, and their training and qualifications vary by grade. The five grades range from the elementary teacher to the highly qualified graduate agrégé, who enjoys the lightest teaching load and the highest prestige and who teaches at the secondary level or higher. The differences have long been a matter of concern, as has the entrenchment of the higher levels of the teaching establishment, which has resisted reforms calling for more uniformity in teacher status as well as changes in method and content orientation, encouraging teacher cooperation, interdisciplinarity, and technological familiarity. Reforms to extend the level of common education, to increase options at the upper secondary level, to strengthen the technological component, and to introduce steps to improve the link between school and work have nonetheless been achieved. Internal reform proposals include the more flexible organization of time and content and the addition of extracurricular activities appropriate to the real life of youth and society. Government forays into decentralization have promoted community links at the school level and school program initiatives. The outcomes will at best affect the system gradually, however.

Other European countries. Most eastern European systems of education follow the Soviet model (see below). In western Europe many countries have been influenced by the British, German, and French systems, but there are numerous variations, some of which are discussed here.

Italy. Education in Italy up to 1923 was governed by the Casati Law, passed in 1859, when the country was being unified. The Casati Law organized the school system on the French plan of centralized control. In 1923 the entire national school system was reformed. The principle of state supremacy was reinforced by introducing at the end of each main course of studies a state examination to be taken by pupils from both public and private schools.

Eight years of schooling has been compulsory since 1948, although this plan was not realized until 1962. The five-year elementary school, for pupils aged six to 11, is followed by the undifferentiated middle or lower secondary school (*scuola media*) for pupils from 11 to 14. There continues to be a strong private (mainly Roman Catholic) interest in preschools and in teacher training for elementary and preschool levels.

Although reform proposals call for an extension of the unitary principle through the five-year upper secondary level, this level is highly diversified, with classical and scientific *licei* (schools) and a vast array of programs in vocational and industrial technical institutes. Shorter courses are given in institutes for elementary teachers and in art schools.

Italian university education

Entrance to Italian universities is gained by successful completion of any of the upper secondary alternatives. Universities are basically the only form of postsecondary education. They require the passing of a variable number of examinations, at the end of which the students sit for a degree (*laurea*), which gives them the title of *dottore.* To be able to exercise any profession, such as that of lawyer, doctor, or business consultant, the student must take a state examination. Students who do not complete their studies in the normal period of time, from four to six years, may remain at the university for several years as *fuori corso* ("out of sequence").

The unification of the lower levels and the expansion of academic and particularly vocational-technical alternatives at the upper level are notable advances, but the Italian education system still suffers from fragmentation and lack of articulation. Indications of low achievement and regional inequalities, in spite of relatively heavy public investment, suggest problems with system effectiveness. The force of conservative political, religious, and educational resistance to change is likely to maintain divisions of policy and outcome.

The Netherlands. The first modern school law in the Netherlands was passed in 1801, when the government laid down the principle that each parish had the right to open and maintain schools. A debate between the proponents of denominational and nondenominational schools went on during the 19th century. The controversy was closed by a law of 1920, which declared that denominational schools were fully equal with state schools, both types being eligible for public funds. The resultant decentralization is unique. Roughly two-thirds of the Dutch school-age children attend private schools. In return for public funds, the private school, which may be Protestant, Roman Catholic, or secular, must provide a curriculum equivalent to that offered by the public schools.

Religious-philosophical diversity is a characteristic feature of Dutch schools. Secondary education comprises four main types, which may be further differentiated: preuniversity, general, vocational, and miscellaneous, which may be part-time. Selection decisions are strongly influenced by examinations. Preprimary and primary schools were recently combined into single eight-year schools for children aged four to 12. Other recent changes include the growth of vocational education at the postsecondary level and the increase in opportunity for females, as indicated by increasing enrollment at higher levels and by the estab-

lishment of special programs, such as that giving women whose schooling was interrupted the chance to return and finish their education.

Switzerland. The Swiss constitution of 1874 provided that each canton or half canton must organize and maintain free and compulsory elementary schools. The federal government exercises no educational function below the university level, except to help finance the municipal and cantonal schools. The Swiss school system thus consists of 26 cantonal systems, each having its own department of education, which sets up its own school regulations. The Swiss Conference of Cantonal Directors of Education has increased its efforts to achieve some educational unity, but great diversity remains.

In general, schooling is compulsory for eight or nine years, beginning at the age of six or seven. The elementary and lower secondary curriculum continues to stress mathematics and language. Cantonal differences in the training of elementary-school teachers remain a matter of concern, but provisions for additional training of in-service teachers are good. Each cantonal system begins to diversify at the lower secondary level and is even further differentiated at the post-compulsory upper secondary level. The pupil's future professional life is a decisive factor in the selection of post-compulsory schooling. Most pupils enter one of the many vocational courses, in which apprenticeship has long played a serious role. Among preuniversity schools, three types have been added to the two traditional ones emphasizing classical languages; the new schools stress mathematics and science (1925), modern languages (1972), and economics (1972). New proposals favour the consolidation of the preuniversity schools.

Swedish educational reforms

Sweden. After World War II the Swedish government began to extend and unify the school system, which had historically been the domain of the Lutheran church. In 1950 the National Board of Education introduced a nine-year compulsory comprehensive school, with differentiation of pupils postponed until late in the program. This *grundskola* replaced all other forms in the compulsory period by 1972–73. Following the unification of the elementary and lower secondary levels was the systematic integration of the upper secondary level, covering ages 16 to 19. This *gymnasieskola* uses organizational and extracurricular means of integration, but students are separated into 25 "lines," of which many are general-academic but most are vocational. Reforms have been implemented to make higher education available to more people, and adult education is encouraged.

The Swedish reform has attracted much attention in Europe for several reasons. It achieved the earliest unequivocal unification of the compulsory-school sector. While moving toward increased levels of integration in the system, the reciprocity of differentiation and integration was used as a principle of school development. As a result, the vocational sector was incorporated into the general upper secondary school. Theory and practice were recognized as components of all programs. The reform process, which specified a long period of experimentation and voluntary action (1950 to 1962) and a correspondingly long period of implementation (1962 to 1972), was singularly well conceived to build planning into participation and practice. The resultant organization is stable but open to change on the same principles. Thus, the new equality thrust goes beyond establishing equal opportunity to providing compensatory measures, even though they sometimes limit free choice, as, for example, in the use of sex quotas to bring women or men into occupations where they had been underrepresented.

Attention has also been focused on the Swedish approach to recurrent education, which introduces the idea of interchanging school and work as early as the secondary level. The coordination of school and work life, which is a worldwide goal, is not only built into institutional programs in Sweden but is also pursued there at a grassroots level through local councils.

(J.A.L./R.L.Sw./R.F.L.)

The United States. As the United States entered the 20th century, the principles that underlie its present educational enterprise were already set. Educational sovereignty rested in the states. Education was free, compulsory, universal, and articulated from kindergarten to university, though the amount of free schooling varied from state to state, as did the age of required school attendance. Although a state could order parents to put their children to their books, it could not compel them to send them to a public school. Parents with sectarian persuasions could send their offspring to religious schools. In principle there was to be equal educational opportunity.

American principle of equal educational opportunity

Expansion of American education. Though such principles remained the basis of America's educational endeavour, that endeavour, like America, has undergone a vast evolution. The once-controversial parochial schools have not only continued to exist but have also increasingly drawn public financial support for programs or students. The currency of privatization, carrying the idea of free choice in a private-sector educational market, strengthens the bargaining position of religious as well as other private schools. The issue of equality has succeeded the issue of religion as the dominant topic of American educational debate. Conditions vary markedly among regions of the country. Definitions of equal opportunity have become more sophisticated, referring increasingly to wealth, region, physical disability, race, sex, or ethnic origin, rather than simply to access. Means for dealing with inequality have become more complex. Since the 1950s, measures to open schools, levels, and programs to minority students have changed from the passive "opportunity" conception to "affirmative action." Measured by high-school completion and college attendance figures, both generally high and continually rising in the United States, and by standardized assessment scores, gains for blacks and other minority students have been noteworthy from the 1970s. Although state departments of education use equalization formulas and interdistrict incentives to reach the poorest areas under their jurisdiction, conditions remain disadvantageous and difficult to address in some areas, particularly the inner cities, where students are mostly minorities. City schools often represent extremes in the array of problems facing youth generally: drug and alcohol abuse, crime, suicide, unwanted pregnancy, and illness; and the complex situation seems intractable. Meeting the needs of a racially and ethnically mixed population has, however, turned from the problem of the cities and from an assimilationist solution toward educational means of knowing and understanding the disadvantaged groups. States have mandated multicultural courses in schools and for teachers. Districts have introduced bilingual instruction and have provided instruction in English as a second language. Books have been revised to better represent the real variety in the population. The status of women has been given attention, particularly through women's studies, through improved access to higher education (women are now a majority of U.S. college students) and to fields previously exclusive to men, and through attempts to revise sexist language in books, instruction, and research.

Programs of special education

The idea persists that in the American democracy everyone, regardless of condition, is expected to have a fair chance. Such is the tenet that underlay the establishment of the free, tax-supported common school and high school. As science pointed the way, the effort to bridge the gulf between the haves and have-nots presently extended to those with physical and mental handicaps. Most states and many cities have long since undertaken programs to teach the handicapped, though financially the going has been difficult. In 1958 Congress appropriated $1 million to help prepare teachers of mentally retarded children. Thenceforward, federal aid for the handicapped steadily increased. With the Education for All Handicapped Children Act of 1975—and with corresponding legislation in states and communities—facilities, program development, teacher preparation, and employment training for the handicapped have advanced more rapidly and comprehensively than in any other period. Current reforms aim to place handicapped children in the least restrictive environment and, where possible, to "mainstream" them in regular schools and classes.

As the century began, American youths attended an eight-year elementary school, whereupon those who continued

went to a four-year high school. This "eight–four system" wholly prevailed until about 1910, when the "six–three–three system" made a modest beginning. Under the rearrangement, the pupil studied six years in the elementary and three in the junior and senior high schools, respectively. Both systems are in use, there being almost the same number of four-year high schools and three–three junior–senior high school arrangements. There has been a change at the elementary–junior high connection to include a system in which children attend an elementary school for four or five years and then a middle school for three or four years. The rapid growth of preschool provisions, with the establishment of an immense body of early-childhood teachers, day-care workers, new "nannies," producers of learning materials, and entrepreneurs, has secured the place of the kindergarten as an educational step for five-year-olds and has made available a wide, but mainly non-public, network of education for younger children.

In 1900 only a handful of the lower school's alumni—some 500,000—advanced into the high school. Of those who took their high-school diploma during this early period, some three out of every four entered college. The ratio reversed, as high-school enrollments swelled 10-fold over the first 50 years of the century, with only one of every four high-school graduates going on to higher learning. As even more students finished high school (more than 75 percent by 1980), demands for access to the post-secondary level increased until nearly half of all high-school graduates, or nearly one-third of the age group, were entering college.

From such experimental programs as the Dalton Plan, the Winnetka Plan, and the Gary Plan, and from the pioneering work of Francis W. Parker and notably John Dewey, which ushered in the "progressive education" of the 1920s and '30s, American schools, curricula, and teacher training have opened up in favour of flexible and cooperative methods pursued within a school seen as a learning community. The attempt to place the nature and experience of the child and the present life of the society at the centre of school activity was to last long after progressive education as a defined movement ended.

Some retrenchment occurred in the 1950s as a result of scientific challenges from the Soviet Union in a period of international political tension. Resulting criticisms of scientific education in the United States were, however, parried by educationists. America's secondary school attuned itself more and more to preparing the young for everyday living. Consequently, though it still served prospective collegians the time-honoured academic fare, it went to great lengths to accommodate the generality of young America with courses in automobile driving, cookery, carpentry, writing, and the like. In addition to changes in the form of earlier practical subjects, the curriculum has responded to social issues by including such subjects as consumer education (or other applications of the economics of a free-enterprise society), ethnic or multicultural education, environmental education, sex and family-life education, and substance-abuse education. Recent interest in vocational-technical education has been directed toward establishing specialized vocational schools, improving career information resources, integrating school and work experience, utilizing community resources, and meeting the needs of the labour market.

Expansion of American high-school curriculum

National prosperity and, even more, the cash value that a secondary diploma was supposed to bestow upon its owner enhanced the high school's growth. So did the fact that more and more states required their young to attend school until their 16th, and sometimes even their 17th, birthday. Recently, however, economic strains, the ineffectiveness of many schools, and troubled school situations in which the safety of children and teachers has been threatened have led to questions about the extension of "compulsory youth" in high schools.

Criticisms have also been leveled at the effects and after-effects on education of 1960s idealism and its conflict with harsh realities. The publicized emphases on alternatives in life-style and on deinstitutionalization were ultimately, in their extreme form, destructive to public education. They were superseded by conservative attitudes favouring a return to the planning and management of a clearly defined curriculum. The dramatic fall in scores on the Scholastic Aptitude Test (a standardized test taken by a large number of high-school graduates) between 1963 and 1982 occasioned a wave of public concern. A series of national, state, and private-agency reviews followed. The report of the National Commission on Excellence in Education, *A Nation at Risk* (1983), set the tone. The emphasis was now on quality of school performance and the relation of schooling to career. The main topics of concern were the curriculum, standardization of achievement, credentialing, and teacher preparation and performance. In order to clarify what is expected of teachers and students, states have increasingly detailed curricula, have set competency standards, have mandated testing, and have augmented the high-school diploma by adding another credential or by using transcripts to show superior achievement. Curriculum reforms have accentuated the academic basics, particularly mathematics, science, and language, as well as the "new basics," including computers. Computers have become increasingly important in education not only as a field of study but also as reference and teaching aids. Teachers are using computers to organize and prepare course materials; children are being taught to use computers at earlier ages; and more and more institutions are using computer-assisted instruction systems, which offer interactive instruction on a one-on-one basis and can be automatically modified to suit the user's level of ability. Other technological developments, such as in broadcasting and video production, are being employed to increase the availability of quality education.

The reports on the state of education also expressed concern for gifted children, who have tended to be neglected in American education. Until psychologists and sociologists started to apply their science to the superior child, gifted children were not suspected of entertaining any particular problems, apart from occasionally being viewed as somewhat freakish. Eventually, however, augmented with federal, state, and sometimes foundation money, one city after another embarked on educational programs for the bright child. From the 1970s, gifted children were directly recruited into special academic high schools and other local programs. American education is still aimed at broadening or raising the level of general provision, however, so neither programs for the gifted nor those for vocational education have been treated as specifically as in some other countries.

Federal involvement in local education

Although the U.S. Constitution has delegated educational authority to the states, which have in turn passed on the responsibility for the daily administration of schools to local districts, there has been no lack of federal counsel and assistance. Actually, national educational aid is older than the Constitution, having been initiated in 1787 in the form of land grants. Seventy-five years later the Morrill Act disbursed many thousands of acres to enable the states to promote a "liberal and practical education." Soon thereafter, the government created the federal Department of Education under the Department of the Interior and, in 1953, established the Office of Education in the Department of Health, Education, and Welfare. As the independent Department of Education from 1980, this agency has taken a vigorous role in stating national positions and in researching questions of overall interest. Its findings have proved influential in both state and local reforms.

Financing of education is shared among local districts, states, and the federal government. Beginning with the Smith–Lever Act of 1914, Congress has legislated measure upon measure to develop vocational education in schools below the college plane. A new trail was opened in 1944, when the lawgivers financed the first "GI Bill of Rights" to enable veterans to continue their education in school or college.

During the 1960s, school difficulties experienced by children from disadvantaged families were traced to lack of opportunities for normal cognitive growth in the early years. The federal government attempted to correct the problem and by the mid-1960s was giving unprecedented funding toward compensatory education programs for disadvantaged preschool children. Compensatory interven-

tion techniques include providing intensive instruction and attempting to restructure home and living conditions. The Economic Opportunity Act of 1964 provided for the establishment of the Head Start program, a total program that was designed to prepare the child for success in public schools and that includes medical, dental, social service, nutritional, and psychological care. Head Start has grown steadily since its inception and has spawned similar programs, including one based in the home and one for elementary-school-age children. In the 1970s child development centres began pilot programs for children aged four and younger. Other general trends of the late 1970s include: extending public schools downward to include kindergarten, nursery school, child development centres, and infant programs; organizing to accommodate culturally different or exceptional children; including educational purposes in day care; extending the hours and curriculum of kindergartens; emphasizing the early-childhood teacher's role in guiding child development; "mainstreaming" handicapped children; and giving parents a voice in policy decisions. Early-childhood philosophy has infiltrated the regular grades of the elementary school. Articulation or interface programs allow preschool children to work together with first graders, sharing instruction. Extended to higher grades, the early-childhood learning methods promote self-pacing, flexibility, and cooperation.

Changes in higher education. The pedagogical experimentalism that marked America's elementary learning during the century's first quarter was less robust in the high school and feebler still in the college. The first venture of any consequence into collegiate progressivism was undertaken in 1921 at Antioch College, in Ohio. Antioch required its students to divide their time between the study of the traditional subjects and the extramural world, for which, every five weeks or so, they forsook the classroom to work at a full-time job. In 1932 Bennington College for women, in Vermont, strode boldly toward progressive ends. Putting a high value on student freedom, self-expression, and creative work, it staffed its faculty largely with successful artists, writers, musicians, and other creative persons, rather than Ph.D.'s. It also granted students a large say in making the rules under which they lived.

Experiments at Antioch, Bennington, Chicago, and Harvard

Such developments in America's higher learning incited gusty blasts from Robert M. Hutchins, president and then chancellor of the University of Chicago from 1929 to 1951. He recommended a mandatory study of grammar, rhetoric, logic, mathematics, and Aristotelian metaphysics. One consummation of the Hutchins prescription is the study of some 100 "great books," wherein reside the unalterable first principles that Hutchins insisted are the same for all men always and everywhere.

The vocationalism that Hutchins deplored was taken to task by several others, but with quite different results—notably by Harvard in its report on *General Education in a Free Society* (1945). Declaring against the high school's heavy vocational leaning, it urged the adoption of a general curriculum in English, science, mathematics, and social science.

In the great expansion of higher education between about 1955 and 1975, when expansionist ideas about curriculum and governance prevailed, colleges became at times almost ungovernable. New colleges and new programs made the higher-education landscape so blurred that prospective students and admissions officers in other countries needed large, coded volumes to characterize individual institutions. The college curriculum, like that of the high school, was altered in response to vocal demands made by groups and had expanded in areas representing realities of contemporary social life. Internal reviews, undergraduate curriculum reforms, and the high standards set by some universities demonstrated to some observers that quality education was being maintained in the university. Other critics, however, felt that grade inflation, the multiplication of graduate programs, and increasing economic strains had led to a decline in quality. Financial problems and conservative reactions to the more extreme reforms led some universities to place a strong emphasis on management.

The junior college

Probably the most significant change in higher education has been the establishment and expansion of the junior college, which was conceived early in the century by William Rainey Harper, president of the University of Chicago. He proposed to separate the four-year college into an upper and a lower half, the one designated as the "university college" and the other as the "academic college." The junior college is sometimes private but commonly public. It began as a two-year school, offering early college work or extensions to secondary education. It has since expanded to include upper vocational schools (including a wide range of technical and clerical occupations), community colleges (offering vocational, school completion, and leisure or interest courses), and pre- or early-college institutions. Junior colleges recruit from a wide population range and tend to be vigorous innovators. Many maintain close relationships with their communities. Colleges limited to the undergraduate level, especially in articulated state systems, may not differ much from well-developed junior colleges.

Professional organizations. American educators began to organize as early as 1743, when the American Philosophical Society was founded, and they have been at it increasingly ever since. Not a few of their organizations, such as the American Historical Association, the Modern Language Association of America, and the American Home Economics Association, are for the advancement of some specialty. Others are more concerned with the interests of the general educational practitioner. Of these the National Education Association (NEA) is the oldest. Founded in 1857, it undertook "to elevate the character and advance the interest of the teaching profession." Despite its high mission, it threw off no sparks, and it was not until after 1870 that it began to grow and prosper. With headquarters in Washington, D.C., the NEA conducts its enormous enterprise through a brigade of commissions and councils. A youngster by comparison, the American Federation of Teachers, an affiliate of the AFL-CIO, was formed in 1916. Through collective bargaining and teachers' strikes, it has successfully obtained for teachers better wages, pensions, sick leaves, academic freedom, and other benefits. The distinction between a union and a professional organization is neither as clear nor as important an issue as it was in earlier days.

Such bodies as the American Association of Colleges for Teacher Education, the American Association of University Professors, the American Educational Research Association, the National Commission on Teacher Education and Professional Standards, and the National Council for Accreditation of Teacher Education have laboured industriously and even with a fair success to bring order and dignity to the teaching profession. Nevertheless, teaching has become an increasingly arduous profession in the United States. Even the security formerly assocated with the profession is in question as waves of teacher shortages and surpluses generate frantic responses by educational authorities. Recent educational reviews have addressed teaching inadequacies by encouraging prospective teachers to earn degrees in other subjects before beginning studies in the field of education. They have recommended establishing proficiency tests, regular staff-development activities, certification stages, and workable teacher-evaluation and dismissal procedures. They insist on the necessity for the reform and evaluation of training programs, and some have questioned the institutional context of teacher training.

(A.E.M./R.F.L.)

Elder members of the British Commonwealth. *Canada.* Although a Canadian nation had been formed by the end of the 19th century, separate political, economic, and geographic influences continued through the 20th century to restrain unified educational development. The historical principle of maintaining minority rights resulted in a truly pluralistic cultural concept, recognized to some extent in religious and linguistic concessions in schools. Each provincial system developed unilaterally, thus producing separately centralized educational units; and, even within a province, the evolving principle of local responsibility and the sparseness of settlement in many areas of Canada challenged the effectiveness of simple control principles. Different production emphases and differential advantages of territorial acquisition after confederation in 1867 cre-

ated basic inequalities among the provinces, with a corresponding effect on schools. Finally, European principles of education were slow to be reconciled with those evolving out of the North American environment. Canadian educational development has consequently been marked by eclectic, pragmatic actions rather than by philosophically or politically unified decisions.

It is nevertheless possible, because of a common national experience and because of the communication stimulated by national development, to describe education in national terms. Educational movements afoot in the early 20th century and associated with "progressive education" (such as child study, kindergarten development, and curriculum integration) had a relatively mild impact on traditional practices and forms. Instruction in the Canadian school remained essentially teacher-centred, with a strong emphasis on obedience and conformity.

Traditional versus progressive education

The major change in school structure occurred at the secondary level. The standard eight-year elementary program was first extended by continuation classes or schools alongside exclusive secondary schools, producing an uneven, overlapping postelementary structure. In the 1930s an expanded school population, reaching into the secondary grades, led to decisive action on compulsory attendance and to standardization of high-school provisions. Junior high schools were introduced in some provinces as a transitional level between elementary and secondary schooling, while some provinces simply developed junior and senior stages of a total secondary program. The two extremes in secondary development were probably represented by Québec and the west. In the French-speaking schools of Québec, the secondary system consisted of private classical colleges leading to a *baccalauréat* on the one side and terminal courses in special schools or institutes on the other. Only after 1956 were public high schools with a variety of courses established. The administration of the system was unified under a ministry of education in 1964, although with continuing provision for local school boards of a distinctly Roman Catholic or Protestant nature. In the western provinces, large regional schools and composite high schools were developed extensively, Alberta having proceeded apace in this direction. British Columbia, following the Chant Commission Report in 1960, reorganized its secondary program to include five core streams, only one of which was academic-technical.

Canadian educational reforms

In general, the secondary curriculum has been modified by expanding the catalog of optional subjects and by reorganizing to include new courses of study. Secondary schools in Canada are now mainly comprehensive and enroll about 85 percent of the age group. After extensive provincial reviews in the 1980s, emphasis has been returned to academic standards and newly placed on the relation of education to work, in response to the economic needs both of society and of the individual. This new emphasis may include teaching specific job skills and industrial information, coordinating vocational and academic studies in school programs, and cooperating with industry through work-study programs. Alternatives to the basic choice between university preparation and a general terminal course have appeared.

In response to the requirements of an expanded school population in the first half of the century and to the later demand for increased access, particularly for women, native Canadians, immigrants, and low-income groups, changes to structure, curriculum, and methods have occurred regularly since the 1960s. Many revisions originated with developments in the United States but took a particularly Canadian form. The first wave of reforms emphasized openness (open-area schools and classrooms, curriculum choice), comprehensiveness (composite high schools, consolidated rural schools, group work, and peer cooperation), and continuity up the school ladder (although with an abundance of alternatives). From the late 1970s, reforms shifted toward renewed emphasis on basic learning, selection of students, moral and social values, increased administrative control, and assessment procedures for school, system, and aggregate student performance.

The educational scene shows characteristics of both periods of reform. Some of the notable innovations include: the provision of preschool classes in most elementary schools or systems; the use in early elementary grades of new educational methods developed at the preschool level; a concentrated attempt to decrease newly discovered functional illiteracy at all levels, including the adult level; the rapid introduction of electronic learning programs and instructional assistance; and direct concern with values instruction, usually secular and oriented to both personal and social issues. Both the attempt to reconcile individual educational requirements with the demands of mass systems and the current emphasis on essential subject matter have led to a search for new techniques of selecting and transmitting knowledge in schools.

The most demanding issues of the second half of the century have reached beyond the traditional time and scope of public schooling: early-childhood education, adult education, private schooling, postsecondary education, and bilingual multicultural provisions. Whether as a reflection of concern over the direction of public schools, of an increasingly pluralistic society, or of affluence among parents, private-school attendance has risen steadily. It is still a small proportion of the school-age group in Canada (less than 5 percent), but the increase in interest as well as in attendance has put pressure on provincial governments for funding. Most provinces now offer limited grants to authorized private schools, though at a level far below public-school financing.

Consistent with Canada's claim to multicultural social development and bolstered by the Canadian Charter of Rights and Freedoms, multicultural and bilingual emphases have made perhaps the strongest single impact on schooling. French-language instruction, both as a mother tongue and as a second language, has expanded in traditionally English-speaking areas. Restrictions have been placed on English-language schooling in Québec as the French-language population struggles for cultural survival in North America. Court challenges against required Christian religious exercises and religious instruction in elementary schools have been successful. Demands have been made to give attention to other languages as languages of instruction and to revise the exclusively Western bias of curriculum content.

Regional patterns of Canadian universities

A new dimension in higher education was added with the establishment of provincial universities in the west (1901–08). This completed a set of regional patterns for university development that has continued to this day. Canadian universities have, within these patterns, drawn their criteria from French, British, or American models. From the 1950s, a boom in Canadian higher education has led to increasingly independent considerations on the role of universities in Canadian development. While the 1950s and '60s saw a great expansion of universities, the 1970s and '80s saw rapid growth in postsecondary, nonuniversity education in provincially funded colleges. These colleges all offer some range of vocational programs. Their relationships with universities vary: some offer university transfer programs (Alberta); some offer university prerequisites (Québec); some have no formal relationship (Ontario). With an increasing student population in a wider range of postsecondary alternatives, the rationalization, planning, and funding of this sector is a primary issue for provincial governments.

The administration of public education is the exclusive responsibility of the provinces, which have worked out schemes of local authority under provincial oversight. Although the specific structure of the departments of education varies among the provinces, they conform to a basic structure. Each is headed by a politically appointed minister of education, who may be advised by a council. The main functions of educational supervision are usually carried out through specific directorates for such areas as curriculum, examinations, vocational education, teacher training and certification, and adult education. Three developments, however, strengthened local autonomy in educational administration. Throughout the second quarter of the 20th century, consolidation of rural schools and administrative units took place in the west, thus resulting in stronger educational units more competent to act independently. Moves toward regional decentralization, especially

in Ontario, Québec, and New Brunswick, produced rather independent subprovincial units. Finally, urban development led to relatively autonomous city school operations. Provincial authority has been reemphasized, however, with the demands for better system articulation and for standardization of requirements, programs, and testing.

Federal influences in Canadian education

Canada's federal government has no constitutional authority in education and therefore maintains no general office dealing directly with educational matters. Federal activities in education are nevertheless carried out under other areas of responsibility, and certain functions of an office of education are subsumed under the secretary of state. The Council of Ministers of Education, Canada, brings together the chief educational officers of the provinces and ensures national communication at governmental level. Under its responsibility for native peoples and its jurisdiction over extra-provincial territories, the federal government, through the Department of Indian Affairs and Northern Development, finances and supervises the education of Indians and Eskimo. In the Yukon, schools are administered by the territorial government, though largely financed from Ottawa.

Through agricultural and technical assistance acts in 1913 and 1919, the federal government began to promote vocational education, and this principle was extended through emergency programs in the depression years of the 1930s and during World War II. More recently, vocational programs of wide scope have been introduced on a principle of federal support and provincial operation. The Technical and Vocational Assistance Act of 1960 was followed by a great surge in vocational education, including the construction of new schools and school additions, special institutes, and the preparation of vocational teachers. Program definitions in this area have become ever broader.

The federal government has maintained and supported the education of armed-forces personnel. Research and development in higher education are promoted directly through grants from national research councils for social sciences and humanities, for the natural sciences and engineering, for medicine, and for the arts. Statistics Canada disseminates organized statistical information on schools and on social factors affecting education. Perhaps less direct but of great importance are national agencies operating in the area of mass communications media, such as the National Film Board. Together, the activities of the federal government not only support but also strongly influence certain areas of education and complete a picture of local-provincial-federal involvement in Canadian education.

Australia. The 20th-century development of Australian education continued to be influenced by British models and to be characterized by the exercise of strong central authority in the states. Yet, because Australian national development has taken place entirely in this century, increasing attention has been given to the role of education in nation building.

Educational systems were built through the establishment of primary schools by the end of the 19th century, the extension of these through continuation programs, and the development of state secondary schools in the early part of the 20th century. The independent secondary schools that offered the bulk of secondary education before 1900 continued to be influential, either as components of the separate Roman Catholic system or as "elite" private schools of denominational or nondenominational character, but the growth of state systems carried the state high schools into numerical prominence.

The early development of educational systems before and around the turn of the century was a crude beginning, the minimal provisions being accentuated by poor teacher preparation, administrative thrift schemes, and excess in the exercise of administrative authority. Improvement of these conditions and systematic positive development can be dated from the Fink Report of 1898 in Victoria and similar reform appeals in other states between 1902 and 1909. The steady pace of progress from that time was broken by a surge of growth and innovation in Australian institutions after World War II.

Education in small, isolated communities throughout the vast Australian area has required special attention. As a means of reaching isolated children and adults, correspondence education was begun in 1914 in Victoria, and other states followed after 1922. The procedures have been gradually refined and the levels extended. More formal early efforts included the introduction of provisional schools, itinerant teachers, and central schools in the outback. The small one-teacher bush schools became typical after federation in 1901. Much attention was given to methods of teaching in the one-room school, earning Australia international recognition for expertise in this area. Progress toward rural school consolidation began in Tasmania in 1936. The Tasmanian model combined special features of school independence, pupil freedom, involvement in agricultural projects, and parental cooperation with the "area school" movement. Recently, there has been a rapid decline in one-teacher all-age schools in Australia in favour of consolidated schools in central locations.

State and federal powers in Australia

Education is a state, rather than a federal, responsibility in Australia. Authority is concentrated in a state department of education. The political head is the minister of education, and the permanent official in charge is the director or director general of education. The main divisions of the department are those for primary, secondary, and technical education, each directed by a senior official; additional divisions, such as for special education or in-service training, are particular to the states. Department policy has been executed through a hierarchy of educational experts. Through the 1980s major changes in administrative organization took place in all state systems toward devolution of authority to local regions and schools. A corporate style of management has become current, using criteria of rationalization, effectiveness, and economic efficiency to guide organizational decisions. Although parties agree on many overall goals, disagreements among state authorities, powerful teachers' unions, and public groups promise the continuation of a politically volatile and changing administrative scene.

Since World War II, with the financial assets of exclusive income-taxing power, the commonwealth (federal) government has played an increasing role in educational development, particularly at the tertiary level. Through the States Grants Act in 1951, the Murray Report in 1957, the Martin Report in 1964, the Karmel Report in 1973, and a series of position papers leading to the 1988 Policy Statement on higher education, the federal government moved into the planning as well as the funding of post-secondary education concurrently with the states. After four decades of rapid expansion in higher education (from less than 50,000 students in 1948 to more than 400,000 in 1988), the government has set a course toward a unified national system at the tertiary level. The government has negotiated directly with higher education institutions, without the traditional buffer of consultative councils, and has moved directly to amalgamate institutions and otherwise to rationalize the system. The organizational rationale is based on the contribution of higher education to the national economic interest, and strategies link higher education to the training needs of the economy. System integrity, efficiency and output measures, and indications of privatization (a private university, tertiary fees, sale of educational services) characterize the political thrust. The Commonwealth Office of Education was established in 1945 to advise on financial assistance to the states and on educational matters generally, to act as a liaison agent among the states and between Australia and other countries, and to provide educational information and statistics. After several title changes, it became, in 1987, the Department of Employment, Education, and Training, bringing together education and training policy with employment strategy at the national level.

About three-quarters of Australian schools are public. The remainder are made up of Roman Catholic schools (which constitute about 80 percent of the nonpublic schools) and other private schools, many of which have considerable influence in the leadership of Australian society. The curriculum and syllabus for each program or course in the state schools is prescribed by the Department of Education, and nonpublic schools generally follow this standard. Since 1965 significant government funding has

been provided to private schools. There has been a resurgence of interest in and a consequent increase of influence from this sector again in recent years.

The Australian school structure

Primary schools are normally of six years' duration, to about age 12, though some schools retain the seventh year of the old pattern. Within primary schools, pupils are organized in grades and advance by annual promotion. Secondary education is offered for five or six years, generally in comprehensive schools. The minimum school-leaving age is 15 (16 in Tasmania). From the 1950s to the mid-1970s, rapid growth occurred throughout the systems, but especially at higher levels. The technical and further education (TAFE) sector has had a singular influence, operating at upper secondary and tertiary levels and providing widespread nonformal activities. TAFE colleges enroll about 700,000 students of school-leaving age annually and serve the great majority of Australian tertiary students. Recent moves have improved cross-crediting between TAFE and other tertiary institutions.

Since the 1970s, three educational goals have emerged: the first emphasizes equality, diversity, devolution, and participation; the second, national and social unity; the third, effective means of managing what had become, because of rapid growth, a huge and nearly ungovernable education sector. As a result, there have been internal reforms in teaching practice, curriculum, school organization, teacher education, and methods of assessment.

The attempts to increase the number of students continuing education and to improve or expand programs to serve the whole population have raised interest in system unification, including such issues as establishing common curricula and stronger Australian content, improving the transition from school to work, and providing equal opportunity for Aborigines, the disabled, and other groups designated as disadvantaged. The government has recently highlighted recognition of the contribution of Aboriginal cultures as well as of Australian studies.

The emphasis on management techniques may conflict with socially broader objectives. The enormous amount of debate current in Australian education has heightened national interest but has hardened ideological lines. The immediacy of political decisions for education and the momentum of present activity will continue to produce system change.

New Zealand. The religious and regional issues that have fettered educational development in other countries of the British Commonwealth were basically settled in New Zealand when the decisions were made in the last quarter of the 19th century to provide wholly secular primary schools and administrative centralization. The major issue in the 20th century has been the achievement of equal educational opportunity. Although New Zealand has accepted the responsibility to educate each child—without racial, social, or narrowly intellectual restriction—to the limits of the child's ability, the unification of the total system to this end has proved quite difficult.

New Zealand's goal of equal educational opportunity

The Education Act of 1914 consolidated the changes that had taken place since 1877. In subsequent reform periods during the '30s and after World War II, barriers to pupil progress through the system were removed or modified. In 1934 the school-leaving certificate examination was established on a broader basis than the university entrance examination, and, in 1936, the proficiency examination governing secondary entrance was abolished. In 1944–45 the school-leaving age was raised to 15; a common core of early secondary studies, including English, social studies, general science, mathematics, physical education, and a craft or fine-arts subject, was established; and universities agreed to accept accredited-school courses without further examination for university entrance. These actions illustrate a gradual but steady facilitation of access through an increasingly coordinated system. The recommendations of the Currie Commission (1962) and the provisions of the Education Act of 1964 continued this direction.

Since 1877 education has been supervised and funded by a central Department of Education, which is headed politically by a minister and permanently by a director general. Administrative duties are generally handled locally, however. Secondary schools are administered by their own boards of governors and primary schools by elected regional boards of education. Universities receive grants negotiated by the University Grants Committee, and grants for other tertiary institutions are administered by the Department of Education. Three regional offices and teams of primary and secondary inspectors link the central Department of Education and the network of local authorities. Education is free until the age of 19 for qualified pupils. University tuition is also paid for successful students.

New Zealand children generally start school at the age of five years and spend eight years in primary school. The secondary system developed through the growth of three separate kinds of school: (1) the district high school, which represented more or less a secondary "top" on a primary school, (2) the independent, academic, one-sex secondary school proper, and (3) the technical school, which took shape between 1900 and 1908. The isolated position of the fee-charging secondary schools of the 19th century was compromised by free-place legislation in 1903, and, by 1914, they were brought into the state system, though retaining a good deal of their independent status. The district high schools remained in the primary system, but their incorporation in the secondary inspection scheme and in secondary teacher classification placed them clearly within that sector of school operation. The technical high school evolved into a general high school with technical bias. Through common departmental inspection, curriculum, and examination standards, and through the effect of the movement for more general postprimary provisions after 1945, the secondary schools increasingly approximated a single pattern.

At the end of the 11th year of schooling, students take the School Certificate examination, a general test that partially determines admittance to the upper secondary level (12th and 13th years). About half the students leave school at this time. Youths qualifying for university entrance find that admission to professional schools is limited. Although the technical institutes and community colleges have been expanded since 1970, demand continues to increase for these programs. Enrollment in teachers' colleges has been limited due to a declining school population.

Private and rural schools

An extensive Roman Catholic private school system grew up after the secularization of state education. From 1970 these schools were subsidized, and after 1975 most became integrated into the state system and funded by the state.

Rural and native education have been given increasing attention in New Zealand. Consolidated schools, served by an extensive transportation system, have long been a feature of rural education. The expansion of community colleges and the establishment of rural education activity programs have extended regional opportunities. Children and adults in isolated districts are served by several correspondence schools. Maori education became a responsibility of the Department of Education in 1879. Since 1962 the government has attempted to balance the need for remediation of deficiencies in general schooling with Maori cultural rights. As in other countries, equity and the relationship between school and work are the two main issues facing the New Zealand school system. Together they represent growing social and economic demands which may not be compatible with the traditional order of schooling.

(R.F.L.)

REVOLUTIONARY PATTERNS OF EDUCATION

Russia: from tsarism to communism. *Before 1917.* At the turn of the 20th century the Russian Empire was in some respects educationally backward. According to the census of 1897, only 24 percent of the population above the age of nine were literate; by 1914 the rate had risen to roughly 40 percent. The large quota of illiteracy reflected the fact that, by this time, only about half the children between eight and 12 attended school. The elementary schools were maintained by the *zemstvo* (local government agencies), the Orthodox church, or the state, the secondary schools mainly by the Ministry of Education.

After the Revolution of 1905 the Duma (parliament) made considerable efforts to introduce compulsory elementary schooling. At the upper stages of the educational

system, progress was significant, too; nevertheless, the secondary schools (*gimnazii, realnyye uchilishcha*) were only to a small degree attended by students of the lower classes, and the higher institutions even less. Preschool education as well as adult education was left to the private initiative of the educationally minded intelligentsia, who were opposed to the authoritarian character of state education in the schools. In 1915–16 the minister of education, Count P.N. Ignatev, started serious reforms to modernize the secondary schools and to establish a system of vocational and technical education, which he regarded as most important for the industrialization of Russia. During the Provisional government (February to October 1917, old style), the universities were granted autonomy, and the non-Russian nationalities received the right of instruction in their native languages. The education system envisaged by the liberal-democratic and moderate Socialist parties was a state common school for all children based on local control and the direct participation of society.

1917–30. After the October Revolution of 1917, the Bolshevik Party proclaimed a radical transformation of education. Guided by the principles of Karl Marx and influenced by the contemporary movement of progressive education in the West as well as in Russia itself, the party and its educational leaders, Nadezhda K. Krupskaya and Anatoly V. Lunacharsky, tried to realize the following revolutionary measures as laid down in the party's program of 1919: (1) the introduction of free and compulsory general and polytechnical education up to the age of 17 within the Unified Labour School, (2) the establishment of a system of preschool education to assist in the emancipation of women, (3) the opening of the universities and other higher institutions to the working people, (4) the expansion of vocational training for persons from the age of 17, and (5) the creation of a system of mass adult education combined with the propaganda of communist ideas. In 1918 the Soviet government had ordered by decree the abolition of religious instruction in favour of atheistic indoctrination, the coeducation of both sexes in all schools, the self-government of students, the abolition of marks and examinations, and the introduction of productive labour. In 1919, special workers' faculties (rabfaks) were created at higher institutions and universities for the development of a new intelligentsia of proletarian descent.

Revolutionary experimentalism

During the period of the New Economic Policy (1921–27), when there was a partial return to capitalistic methods, the revolutionary spirit somewhat diminished, and the educational policy of party and state concentrated on the practical problems of elementary schooling, the struggle against juvenile delinquency, and the schooling of adult illiterates. When the policy of five-year plans began in 1928 under the slogan of "offensive on the cultural front" and with the help of the Komsomol (the communist youth league), the campaign against illiteracy and for compulsory elementary schooling reached its climax.

The Stalinist years, 1931–53. In connection with the policy of rapid industrialization and collectivization of farmers and with the concentration of political power in the hands of Joseph Stalin, the Soviet educational policy in the 1930s experienced remarkable changes. Starting with the decree of 1931, the structure and the contents of school education underwent the following process of "stabilization" in the next few years: (1) four years was laid down as the compulsory minimum of schooling for the rural districts, and seven years for the cities; (2) the new system of general education embraced the grades one to four (*nachalnaya shkola*), the grades five to seven, which continued the elementary stage on the lower secondary level (*nepolnaya srednyaya shkola*), and the grades eight to 10, which provided a full secondary education (*polnaya srednyaya shkola*); (3) the new curriculum was to provide the students with a firm knowledge of the basic academic subjects and was to be controlled by a system of marks and examinations; (4) the decisive role of the teacher within the educational process was reestablished, while the Pioneers and Komsomol organizations (for youth aged 10 to 15 years and 14 to 26 years, respectively) were above all to instill a sense of discipline and an eagerness for learning; (5) manual work disappeared from the school curriculum as well as from the teacher-training institutions. In addition, the ideas of progressive education were rejected, and older Russian traditions began to be cultivated. During World War II the idea of Soviet patriotism emerged fully, penetrating the theory and practice of education. The principles of the outstanding educator Anton S. Makarenko, with their emphasis on collectivism, gained ground upon the former influence of Western educational thought.

The institutions of higher learning were reshaped in the 1930s, too. The number of students in institutions providing secondary specialized education, usually called *tekhnikumy,* rapidly grew from one million in 1927–28 to 3.8 million in 1940–41. The number of students in institutions of higher education (*vyssheye uchebnoye zavedeniye*) grew from 168,554 to 811,700 in the same period. The main characteristics of higher education that developed in this period remained unchanged for the next decades: the paramount task of higher learning was to provide specialized vocational training within the framework of manpower policy and economic plans; strict control of the student's program was to be imposed by the central authorities; and the system of evening and correspondence instruction on the level of higher and secondary specialized education (*vecherneye i zaochnoye obrazovaniye*) was to parallel full-time studies.

During the 1940s, "labour reserve" trade schools and factory schools for skilled and semiskilled labour were filled by drafting youths between the ages of 14 and 17. In the period 1940 to 1958, an average of 570,000 persons were annually subjected to such recruitment. The draft first affected those students who were unsuccessful academically in regular secondary schools and could not achieve even the seventh grade. For youngsters of this kind and for people who could not continue general secondary education, schools for the working youth (*shkoly rabochey molodyozhi*) and schools for rural youth (*shkoly selskoy molodyozhi*) were established in 1943–44 as part-time institutions. The main features of education policy, developed in the late 1930s, remained in force after the war: the orientation of all kinds of schooling and training to the paramount necessities of the economic system; the inculcation of communist discipline and Soviet patriotic attitudes; and finally a rigid control of the whole educational system by party and state administration.

Labour training

The Khrushchev reforms. After the death of Stalin in 1953, changes in official policy affected both education and science. The 20th Party Congress in 1956 paved the way for a period of reforms inaugurated by Nikita S. Khrushchev. The central idea was formulated as "strengthening ties between school and life" at all levels of the educational system. The Soviet reform influenced to a high degree similar reforms in the eastern European countries.

The old idea of polytechnical education was revived, but mainly in the sense of preparing secondary-school students for specialized vocational work in industry or agriculture. Since the early 1950s there had been a growing imbalance between the output of secondary-school graduates desiring higher education and the economic demands of skilled manpower at different levels. The educational reforms of 1958 pursued the aim of combining general and polytechnical education with vocational training in a way that directed the bulk of young people after the age of 15 straight into "production."

The new structure of the school system after 1958 developed as follows: (1) the basic school with compulsory education became the eight-year general and polytechnical labour school, for ages seven to 15 (*vosmiletnyaya shkola*); and (2) secondary education, embracing grades nine to 11, was provided alternatively by secondary general and polytechnical labour schools with production training (*srednyaya obshcheobrazovatelnaya trudovaya politekhnicheskaya shkola s proizvodstvennym obucheniem*) or by evening or alternating-shift secondary general education schools (*vechernyaya smennaya srednyaya obshcheobrazovatelnaya shkola*).

The connection of study and productive work was to be continued during the course of higher education. Great emphasis was laid upon the further expansion of evening and correspondence education both at the level of sec-

ondary specialized education and at the level of the universities and other higher institutes. In the academic year 1967–68, 56.3 percent of all Soviet students in higher education (of the total number of 4,311,000) carried out their studies in this way.

The reform of 1958 also brought a transformation of the former labour-reserve schools into urban vocational-technical schools or rural schools of the same type (*gorodskiye i selskye professionalno-tekhnicheskiye uchilishcha*). As a rule these schools require the completion of the eight-year school, but in fact there are many pupils with lower achievements; the length of training is from one to three years, depending upon the type of career.

Collective education

Besides introducing polytechnical education and productive labour, the Khrushchev reforms emphasized the idea of collective education from early childhood. Preschool education for the age group up to seven years was to be rapidly developed within the newly organized unified crèches and nursery schools (*yasli i detskiye sady*); and, as a new type of education, boarding schools (*shkoly-internaty*) that embraced grades one to eight or one to 11 had been created in 1956. Some party circles wanted this kind of boarding education for the majority of all young people, but development lagged behind planning, and the idea of full boarding education was later abandoned.

The polytechnization of the Soviet school system as it took shape during the Khrushchev period turned out, in the course of its realization, to be a failure. A revision of the school reform was carried out between August 1964 and November 1966 that brought about several important results: (1) the grade 11 of the secondary school (except for the evening school) was abolished; general education returned to the 10-year program; (2) vocational training in the upper grades was retained only in a small number of well-equipped secondary schools; and (3) a new curriculum and new syllabi for all subjects were elaborated. After 1958 hundreds of secondary schools for gifted pupils in mathematics, science, or foreign languages were developed, besides the well-known special schools for music, the arts, and sports. They recruit students mainly from the urban intelligentsia and have therefore sometimes been criticized by adherents of egalitarian principles in education.

From Brezhnev to Gorbachev. Leonid I. Brezhnev assumed leadership after Khrushchev retired in 1964. On Nov. 10, 1966, a decree was issued outlining the new policy in the field of general secondary education. A union–republic Ministry of Public Education was established to augment the already existing central agencies for higher and secondary specialized education and for vocational-technical training. The main aim of educational policy in the 1970s was to achieve universal 10-year education. In 1977 it was claimed that about 97 percent of the pupils who graduated from the basic eight-year school continued their education at the secondary level. An important step toward the realization of universal secondary education was the creation of secondary vocational-technical schools (*srednye professionalno-tekhnicheskiye uchilishcha*) in 1969. These schools offered a full academic program as well as vocational training. Preschool education for children under seven years of age was extended: enrollments in nursery schools, kindergartens, and combined nursery–kindergarten facilities increased from 9.3 million in 1970 to 15.5 million in 1983. The number of institutions for higher education also grew steadily (from 805 in 1970 to 890 in 1983), meeting regional demands. Day, evening, and correspondence courses were provided.

The expansion of secondary education

The quantitative gains achieved during this period were not matched by corresponding improvements in the quality of education. Government authorities, as well as teachers and parents, expressed growing dissatisfaction with student achievement and with student attitude and behaviour. The youngsters themselves often felt alienated from the official value system in education. Furthermore, there was a growing imbalance between the careers preferred by general-school graduates and the national economic requirements for skilled manpower—an unforeseen result of the policy of universal secondary education. Therefore, from 1977 the scope of labour training in the upper grades of the general school was enhanced in order to provide youngsters with a basic practical training and to direct them into so-called mass occupations after leaving school.

In 1984, two years after Brezhnev's death, new reforms of general and vocational education were instituted. Teachers' salaries, which had been lower than other professional incomes, were raised. The age at which children enter primary school was lowered from seven to six years, thus extending the complete course of general-secondary schooling from 10 to 11 years. Vocational training in the upper grades of the general school was reinforced. To meet the requirements of computer literacy, appropriate courses were introduced into the curricula of the general school, even though most schools lacked sufficient equipment. The main emphasis, however, was placed on the development of a new integrated secondary vocational-technical school that would overcome the traditional barriers between general and vocational education.

Perestroika and education. The 1984 reform of Soviet education was surpassed by the course of economic and structural reforms (*perestroika*) instituted since 1986 under the leadership of Mikhail S. Gorbachev. In February 1988 some earlier reforms were revoked, including the compulsory vocational training in the general school and the plans to create the integrated secondary school. Universal youth education was limited to a nine-year program of "basic education," with subsequent secondary education divided into various academic and vocational tracks. The newly established State Committee of Public Education incorporated the three formerly independent administration systems for general schooling, vocational training, and higher education. Even more important was the rise of an educational reform movement led by educationists who favoured an "education of cooperation" (*pedagogika sotrudnichestva*) over the authoritarian and dogmatic principles of collective education that originated in the Stalin period. These theorists advocated individualizing the learning process, emphasizing creativity, making teaching programs and curricula more flexible, encouraging teacher and student participation, and introducing varying degrees of self-government in schools and universities as a part of the proclaimed "democratization" of Soviet society. Some of the proposals were approved by the State Committee; for example, the universities and other institutions of higher learning were granted some autonomy. Other proposals were being tested by teachers in experimental groups.

Education in the non-Russian republics

In the non-Russian republics the language of instruction is a key issue. After the October Revolution of 1917, education in native languages was promoted. In the 1970s, however, the number of Russian-language and bilingual schools grew steadily at the expense of schools offering instruction in the native languages, even in territories with a majority of non-Russian ethnic groups. This Russianization provoked increasing opposition, and in 1987 the Baltic republics (Estonia, Latvia, Lithuania) began to demand the national-cultural autonomy that is formally guaranteed by the Soviet constitution. Political and educational concessions made to the Union republics seemed to herald further decentralization of the so-far uniform and centralized system of education. (O.A.)

China: from Confucianism to communism. *The modernization movement.* The political and cultural decline of the Manchu dynasty was already evident before the 19th century, when mounting popular discontent crystallized into open revolts, the best known of which was the Taiping Rebellion (1850–64). The dynasty's weakness was further exposed by its inability to cope with the aggressive Western powers during the 19th century. After the military defeats administered by the Western powers, even Chinese leaders who were not in favour of overthrowing the Manchus became convinced that change and reform were necessary.

Most of the proposals for reform provided for changes in the educational system. New schools began to appear. Missionary schools led the way in the introduction of the "new learning," teaching foreign languages and knowledge about foreign countries. New schools established by the government fell under two categories: foreign-language schools to produce interpreters and translators and schools

for military defense. Notable among the latter were the Foochow Navy Yard School to teach shipbuilding and navigation and a number of academies to teach naval and military sciences and tactics.

China's defeat by Japan in 1894–95 gave impetus to the reform movement. A young progressive-minded emperor, Kuang-Hsü, who was accessible to liberal reformers, decided upon a fairly comprehensive program of reform, including reorganizing the army and navy, broadening the civil service examinations, establishing an Imperial University in the national capital and modern schools in the provinces, and so on. The imperial edicts in the summer of 1898 spelled out a program that has been called the Hundred Days of Reform. Unfortunately for China and for the Manchu dynasty, conservative opposition was supported by the empress dowager Tz'u-hsi, who took prompt and peremptory action to stop the reform movement. The edicts of the summer were reversed and the reforms nullified. Frustration and disappointment in the country led in 1900 to the emotional outburst of the Boxer Rebellion.

Hundred Days of Reform and the aftermath

After the Boxer settlement even the empress dowager had to accept the necessity of change. Belatedly, she now ordered that modern schools, teaching modern subjects such as Western history, politics, science, and technology, along with Chinese classics, be established on all levels. The civil service examinations were to be broadened to include Western subjects. A plan was ordered to send students abroad for study and recruit them for government service upon return from abroad. But these measures were not enough to meet the pressing demands now being presented with increasing forcefulness. Finally, an edict in 1905 abolished the examination system that had dominated Chinese education for centuries. The way was now cleared for the establishment of a modern school system.

The first modern school system was adopted in 1903. The system followed the pattern of the Japanese schools, which in turn had borrowed from Germany. Later, however, after establishment of the republic, Chinese leaders felt that the Prussian-style Japanese education could no longer satisfy the aspirations of the republican era, and they turned to American schools for a model. A new system adopted in 1911 was similar to what was then in vogue in the United States. It provided for an eight-year elementary school, a four-year secondary school, and a four-year college. Another revision was made in 1922, which again reflected American influence. Elementary education was reduced to six years, and secondary education was divided into two three-year levels.

Education in the republic. The first decade of the republic, up to the 1920s, was marked by high hopes and lofty aspirations that remained unfulfilled in the inclement climate of political weakness, uncertainty, and turmoil. The change from a monarchy to a republic was too radical and too sudden for a nation lacking any experience in political participation. The young republic was torn by political intrigue and by internecine warfare among warlords. There was no stable government.

A school system was in existence, but it received scant attention or support from those responsible for government. School buildings were in disrepair, libraries and laboratory equipment were neglected, and teachers' salaries were pitifully low and usually in arrears.

New intellectual movements during the republic

It was, nevertheless, a period of intellectual ferment. The intellectual energies were channeled into a few movements of great significance. The first was the New Culture Movement, or what some Western writers have called the Chinese Renaissance. It was at once a cordial reception to new ideas from abroad and a bold attempt to reappraise China's cultural heritage in the light of modern knowledge and scholarship. China's intellectuals opened their minds and hearts to ideas and systems of thought from all parts of the world. They eagerly read translated works of Western educators, philosophers, and literary writers. There was a mushroom growth of journals, school publications, literary magazines, and periodicals expounding new ideas. It was at this time that Marxism was introduced into China.

Another movement of great significance was the Literary Revolution. Its most important aspect was a rebellion against the classical style of writing and the advocacy of a vernacular written language. The classics, textbooks, and other respectable writings had been in the classical written language, which, though using the same written characters, was so different from the spoken language that a pupil could learn to read without understanding the meaning of the words. Now, progressive scholars rejected the heretofore respected classical writing and declared their determination to write as they spoke. The new vernacular writing, known as *pai-hua* ("plain speech"), won immediate popularity. Breaking away from the limitations of stilted language and belaboured forms, the *pai-hua* movement was a boon to the freedom and creativity released by the New Thought Movement and produced a new literature attuned to the realities of contemporary life.

A third movement growing out of the intellectual freedom of this period was the Chinese Student Movement, or what is known as the May Fourth Movement. The name of the movement rose from nationwide student demonstrations on May 4, 1919, in protest against the decision of the Paris Peace Conference to accede to the Japanese demand for territorial and economic advantages in China. So forceful were the student protests and such overwhelming support did they get from the public that the weak and inept government was emboldened to take a stand at the conference and refused to sign the Versailles Treaty. The students thus had a direct hand in changing the course of history at a crucial time, and from now on Chinese students constituted an active force on the political and social scene.

Education under the Nationalist government. Nationalist China rose in the mid-1920s amid a resurgence of nationalism and national consciousness stimulated by post-World War I developments. It was led by the Kuomintang, the political party organized by Sun Yat-sen, the founder of the republic. Cognizant of the popular appeal of nationalism, the Kuomintang set up a Nationalist government pledged to achieve national unity at home and national independence from foreign control as prerequisites to a program of modernization and national reconstruction. In education, it set out to systematize and stabilize a shaky and ill-supported school system and use it as a means of national regeneration. Schools were assured of financial support, however inadequate, and placed under strict supervision and firm control by public authorities.

State control and centralization of education

State control of education by means of centralized administration was instituted. Measures were adopted to correct the abuses and chaos that had resulted from the laissez-faire educational policy of the warlords. Decrees and regulations issued by the Nationalist Ministry of Education were strictly enforced, with the aid of a centrally administered system of inspection and accreditation. Detailed regulations covered the curricula of schools on all levels, minimum standards of achievement, teaching procedures, teachers' qualifications, and specifications for school buildings, libraries, laboratories, and the like. Private schools were permitted but were as subject to government control as public schools and were required to follow the same regulations with regard to curriculum and all other details.

A uniform system of schools was in effect throughout the country. Elementary education was provided in the four-year primary school, followed by the two-year higher elementary. In areas where there were not enough funds to support longer courses, there were abbreviated schools having only one or two grades. Theoretically, the government was committed to the goal of four-year compulsory education, but financial problems prevented an early realization of this goal. Adult education was given much attention in adult schools, in mass education projects, and in different forms of "social education." The latter term encompassed a variety of educational agencies outside the schools, such as libraries, museums, public reading rooms, recreational centres, music, sports, radio broadcasting, and films. Reduction of illiteracy was a major objective.

There were three parallel types of secondary education: the academic middle school, the normal school, and the vocational school. To counteract the traditional preference for the academic type of education, the government restricted the growth of the academic middle school. At the same time, vocational schools were encouraged.

Nationalist promotion of "practical studies"

A major objective of government policy was to promote "practical studies." In secondary education, "practical studies" meant the development of vocational and technical schools and more attention to science and laboratory experience in middle schools. In higher education, measures were taken to steer students away from liberal arts, law, education, and commerce to the "practical courses" of science, engineering, technology, agriculture, and medicine. Government grants for private as well as public colleges were usually designated for the science program. As a result of this policy, the years prior to World War II saw a steady increase of enrollment in the "practical courses" of study and a corresponding decline of enrollment in the arts–law–education–commerce courses. The increase of interest in science was also evident in the secondary schools.

It may be said that the thrust of educational policy in Nationalist China was to rectify the imbalance of the past, especially the nonvocational literary tradition of premodern days. In the attempt to counteract past tendencies, however, it was possible that the pendulum might swing to the other extreme. Some educators expressed the fear that the promotion of "practical studies" might lead to a narrow, utilitarian concept of education and a neglect of the humanities and social sciences. Others were uneasy over the danger of regimentation through centralized administration. Nevertheless, education under the Nationalist government did succeed in establishing an effective national system of education, promoting science and technical studies and correcting the abuses and irregularities of the earlier period. Thanks to dependable financial support, state schools and universities gained in prestige and academic performance until they were recognized as among the outstanding educational institutions of the country.

Other accomplishments of this period include the growth of postgraduate education and research, the general acceptance of coeducation in elementary and higher education, and the use of the Kuo yü (National Tongue) as an effective means of unifying the spoken language and thus overcoming the difficulties of local dialects.

Education under communism. The communist revolution aimed at being total revolution, demanding no less than the establishing of a new society radically different from what the orthodox communists called the feudal society of traditional China. This new society called for people with new loyalties, new motivations, and new concepts of individual and group life. Education was recognized as playing a strategic role in achieving this revolution and development. Specifically, education was called upon to produce, on the one hand, zealous revolutionaries ready to rebel against the old society and fight to establish a new order and, at the same time, to bring up a new generation of skilled workers and technical personnel to take up the multitudinous tasks of development and modernization.

The People's Republic of China generally makes no distinction between education and propaganda or indoctrination. All three share the common task of changing man. The agencies of education, indoctrination, and propaganda are legion—newspapers, posters, and propaganda leaflets, neighbourhood gatherings for the study of current events, as well as political rallies, parades, and many forms of "mass campaigns" under careful direction. It is evident that the schools constitute only a small part of the educational program.

When the Communists came to power in 1949, they took up three educational tasks of major importance: (1) teaching many illiterate people to read and write, (2) training the personnel needed to carry on the work of political organization, agricultural and industrial production, and economic reform, and (3) remolding the behaviour, emotions, attitudes, and outlook of the people. Millions of cadres were given intensive training to carry out specific programs; there were cadres for the enforcement of the agrarian law, the marriage law, the electoral law; some were trained for industry or agriculture, others for the schools, and so on. This method of short-term ad hoc training is characteristic of communist education in general.

Because the new Communist leaders had no experience in government administration, they turned to their ideological ally, the Soviet Union, for aid and guidance. Soviet advisers responded quickly, and Chinese education and culture, which had been Westernized under the Nationalists, became Sovietized. An extensive propaganda campaign flooded the country with hyperbolic eulogies of Soviet achievements in culture and education. The emphasis on Soviet cultural supremacy was accompanied by the repudiation of all Western influence.

Soviet influence on Chinese education

A major agency designed to popularize the Soviet model was the Sino-Soviet Friendship Association (SSFA), inaugurated in October 1949 immediately after the new regime was proclaimed. Headed by no less a personage than Liu Shaoqi, the second highest Chinese Communist leader, the association extended its activities to all parts of the country, with branch organizations in schools, factories, business enterprises, and government offices. In schools, students were urged to enlist as members of the association and to participate in its activities. In many schools more than 90 percent of the students became SSFA members. Throughout the nation, the SSFA sponsored exhibits, motion pictures, mass meetings, parades, and lectures to engender interest in the Soviet Union and in the study of Russian language, education, and culture.

Soviet advisers drew up a plan for the merging and geographic redistribution of colleges and universities and for the reorganization of collegiate departments and areas of specialization in line with Soviet concepts. Colleges and departments of long standing were eliminated without regard to established traditions or to the interests and scholarly contributions of their faculties. Russian replaced English as the most important foreign language.

From curriculum content to teaching methods, from the grading system to academic degrees, communist China followed the Soviet model under the tutelage of Soviet advisers, whose wisdom few dared question. Even the new youth organizations (which displaced the Boy Scouts and Girls Scouts) were comparable to the Pioneers and Komsomols of the U.S.S.R. According to one report, at the peak of the Sovietization frenzy, the first lesson in a Chinese-language textbook used in primary schools was a translation from a Russian textbook.

Never before in the history of education in China had such an extensive effort been made to imitate the education of a foreign country on such a large scale within such a short period of time. Nevertheless, there were many reasons why the campaign did not produce many lasting changes in Chinese education. Russian education and culture had not been well known in China, and the nation was not psychologically prepared for such a sudden and intensive dose of indoctrination to "learn from the Soviet Union." Students, teachers, and intellectuals in general, who would have reacted favourably to a reform to make education more Chinese, were skeptical of the wisdom of switching from Western influence to Soviet influence.

Ideology and education

Chinese leaders justified the indiscriminate imitation of the Soviet model on ideological grounds. The Soviet Union was the leader of the socialist countries; Lenin and Stalin were the shining lights that led the people of the world in their struggle for freedom and equality; the supremacy of the Soviet Union had proved the superiority of socialism over capitalism.

The paramount importance of ideology in education may also be seen in other ways. Ideological and political indoctrination was indispensable to all levels of schools and to adult education and all forms of "spare-time education." It consisted of learning basic tenets of Marxism–Leninism and studying documents describing the structure and objectives of the new government as well as major speeches and utterances of the party and government leaders. Its aim was to engender enthusiasm for the proletarian-socialist revolution and fervent support for the new regime. Class and class struggle were related concepts that occupied a central place in the ideology, and a specific aim of education was to develop class consciousness so that all citizens, young and old, would become valiant fighters in the class struggle. School regulations stipulated that 10 percent of the curriculum should be set aside for ideological and political study, but, in practice, ideology and politics were taught and studied in many other sub-

jects, such as language, arithmetic, and history. Ideology and politics permeated the entire curriculum and school life, completely dominating extracurricular activities.

Among the most important educational changes of this period was the establishment of "spare-time" schools and other special schools for peasants, workers, and their families. Adults attended the spare-time school after their day's work or during the lax agricultural season. Workers and peasants were admitted to these schools by virtue of their class origin. Political fervour and ideological orthodoxy replaced academic qualifications as prerequisites for further study. As a result of the Cultural Revolution of 1966–76, higher education was greatly curtailed and production and labour were emphasized. Mao Zedong, the Communist Party chairman, issued a directive sending millions of students and intellectuals into the rural areas for long-term settlement and "reeducation." He asserted that the intelligentsia could overcome the harmful effects of bourgeois-dominated education only by identifying with the labouring masses through engaging in agricultural and industrial production. Proletarian leadership was also emphasized, as "Mao Zedong thought propaganda teams," made up of workers, peasants, and soldiers who were well-versed in quotations from Chairman Mao but otherwise often barely literate, took over the management of almost all educational institutions.

Post-Mao education. After Mao's death on Sept. 9, 1976, the new leaders lost no time in announcing a turnabout of ideological-political emphasis from revolution to development. They decreed that all effort should be directed toward "the four modernizations" (industry, agriculture, national defense, and science and technology). The primary task of education was to train the personnel needed to speed up the modernization program.

The post-Mao schools are very different from those of the revolutionary education. The conventional school system has been reinstated. Full-time schools have again become the mainstay of a system of coordinated schools, with orderly advance from level to level regulated by examinations. School discipline has been restored, and due respect for teachers is expected of students. Serious study is not to be overshadowed by extracurricular activities; the line of demarcation between formal and informal education is clearly drawn. The main task of students, said the Communist leader Deng Xiaoping, is "to study, to learn book knowledge," and the task of the school is to make "strict demands on students in their study . . . making such studies their main pursuit."

Acquisition of knowledge is again a legitimate aim of education. Academic learning and the development of the intellect have returned after a decade of banishment. Efforts are being made to raise academic standards not only in the universities but in the lower schools as well.

Key schools

The "key schools," outstanding schools that elevate the standards of teaching and learning and serve as models for others, have been revived. They are provided with funds for well-equipped libraries and laboratories and are staffed with highly qualified teachers. Condemned during the Cultural Revolution as "little treasure pagodas" that catered to bourgeois children to the exclusion of workers, peasants, and soldiers, these centres of academic scholarship are now hailed as the standard-bearers of quality education.

Examinations have returned with a vengeance. Every year the government sets a date and time for the unified competitive college examination. High-school graduates take the examination locally, indicating, in order of preference, the colleges they would like to attend if they pass.

Although in theory every college has a president, a vice president, deans, and the like, the real educational policymaker is the Communist Party organization in each school. School presidents or other administrators must often be party members, but even they cannot make decisions without the full cooperation of party representatives. Recently there have been demands for reforms giving more power to school administrators and faculty members.

Despite the renewed emphasis on academics, the national budget allotment for education is insufficient. Administrators and faculty members are underpaid. The government thus permits teachers to have secondary occupations.

Communism and the intellectuals. Throughout China's long history, the intellectuals considered themselves the preservers and transmitters of the precious culture of their country. Their road to success was not always smooth, but the intellectuals were strengthened by the belief that once they won recognition as first-rank scholars they would be rewarded with position, honour, and lasting fame.

The attitude of the Chinese communists toward intellectuals is in large measure influenced by their ideology. While workers and peasants were raised to the top position, the intellectuals were downgraded because they were considered products of bourgeois and feudal education and perpetuators of bourgeois ideology. The communist policy was to "absorb and reform" the intellectuals.

The intellectuals were made to undergo thorough thought remodeling to be "cleansed" of bourgeois ideas and attitudes. The remodeling began with relatively mild measures, such as "political study" and "reeducation." The policy became increasingly oppressive in the 1950s when intellectuals were pressured to take part in the class struggle of the land reform and in orchestrated attacks on university professors, writers, artists, and intellectuals in different walks of life. The intellectuals, especially those who had studied in Western schools or had been employed by Western firms, were forced to write autobiographies giving details of their reactionary family and educational background, pinpointing their ideological shortcomings, and confessing their failings.

Following Khrushchev's 1956 speech criticizing Stalin, violence broke out in Poland and Hungary. This worried Mao, who agreed to try Premier Zhou Enlai's proposal to relax the Communist Party's pressure on intellectuals. This resulted in the slogan "Let a hundred flowers bloom, a hundred schools of thought contend." Mao indicated that intellectuals would be allowed to speak freely.

The Hundred Flowers Campaign

The result, however, was unexpected and shocking. Once they began to speak freely, the intellectuals unleashed a torrent of angry words, fierce criticisms, and open attacks upon the repressive measures under which they had suffered. Some recanted the confessions they had made under duress; others went so far as to denounce the Communist Party and its government. To avoid a more serious outburst of explosive ideas and emotions, the government decided to put a stop to the "blooming–contending." Outspoken critics were labeled rightists, and an anti-rightist campaign not only silenced the intellectuals but also placed them under more restrictive controls than before. The "flowers" wilted and the "schools" were muffled.

During the Cultural Revolution, Mao's criticism of the intellectuals instigated young radicals all over the country to join the struggle against the intellectuals. Students were urged to slap and to spit at their teachers; insult, humiliation, and torture were common. Some teachers chose suicide. Others were sent to May 7th cadre schools or to the countryside to be reformed by labour.

After Mao's death and the repudiation of the radical extremists, the intellectuals began to grow stronger. A movement called "Peking Spring" was launched in November 1978. Huge wall-posters condemning the communist regime appeared on Peking's so-called Democracy Wall. The movement's leaders expanded the modernization program by adding a fifth modernization which clearly emphasized democracy, freedom, and human rights. The "Peking Spring" movement was short-lived, but Chinese intellectuals in the United States and Hong Kong, as well as in China, continued to organize themselves and to advocate democracy and freedom. In China Fang Lizhi, an astrophysicist, toured university campuses speaking against the repression that he believed had killed the initiative and creativity of Chinese scholars. In the spring of 1989 a grand prodemocracy demonstration took place in Peking. The university students took the lead, demanding a higher allotment of funds for education and protesting corruption, but people from all walks of life joined the demonstration. The movement drew attention and support both at home and abroad; but it was soon suppressed by the government, and the country, including educational affairs, continues to be controlled by the Communist Party. (T.H.C.)

PATTERNS OF EDUCATION IN NON-WESTERN OR DEVELOPING NATIONS

Japan. *Education at the beginning of the century.* Between 1894 and 1905 Japan experienced two conflicts, the Sino-Japanese and Russo-Japanese wars, which increased nationalistic feelings; it also experienced accelerated modernization and industrialization. In accord with the government's new nationalism and efforts to modernize the country, educational reform was sought. The Japanese education system took as its model the western European educational systems, especially that of Germany. But the basic ideology of education remained the traditional one outlined in 1890 in the Imperial Rescript on Education (Kyōiku Chokugo).

Effect of Japan's nationalism and economic growth

In 1900 the period of ordinary elementary schooling was set at four years, and schooling was made compulsory for all children. At the same time, the cost of compulsory education was subsidized from the national treasury. In 1907 the period of compulsory education was extended from four to six years. As the educational system gradually improved and as modernization progressed and the standard of living increased, school enrollments soared. The percentage of elementary-age children in school rose from 49 in 1890 to 98 in 1910.

In those days, boys and girls in primary school studied under the same roof, though in separate classrooms. In secondary education, however, there were entirely separate schools for boys and girls—the *chūgakkō,* or middle school, for boys and the *jogakkō,* or girls' high school, both aiming at providing a general education. Other than these, there was the *jitsugyōgakkō,* or vocational school, which was designed to afford vocational or industrial education for both boys and girls. All three secondary schools were for students who had completed the six- or four-year course of primary education.

As for the elementary and secondary curriculum, the Imperial Rescript on Education made it clear that traditional Confucian and Shintō values were to serve as the basis of moral education. This emphasis was implemented by courses on "national moral education" (*shūshin*), which served as the core of the curriculum. In 1903 a system of national textbooks was enacted, giving the Ministry of Education the authority to alter texts in accordance with political currents.

To meet the demand for an expansion of education, a new system for training primary-school teachers was established under the Normal School Order of 1886 and subsequently developed under the strong control of the government. All the normal schools were run by the prefectures, and none was private. At first only the graduates of the higher primary schools were qualified for the normal school, but in 1907 a new course was introduced for graduates of the middle schools and the girls' high schools. For training secondary-school teachers, there was after 1886 the *kōtō Shihangakkō,* or higher normal school for women. Additionally, temporary teachers' training institutes were established after 1902. These were all state-run. There were also state-run institutes for training vocational-school teachers.

Japan's imperial universities

For higher education, there were academies for the study of Confucianism, but a university of the European variety did not appear in Japan until 1877. In that year the University of Tokyo was founded, with four faculties—law, physical sciences, literature, and medicine. In the early years, research and education were dominated by foreigners: most programs were taught in the English language by English and American teachers or, in the medical faculty, in the German language by German instructors. In 1886 the University of Tokyo was renamed the Imperial University by imperial order and, as a state institution, was assigned to engage exclusively in research and instruction of such sciences and technology as were considered useful to the state. Modern Western sciences formed the core of this research and instruction, though some traditional Japanese learning was revived. Engineering and agricultural science were added to the four established faculties. Tokyo Imperial University borrowed much of the style and mode of the German universities and served as the model for the imperial universities established thereafter. Meanwhile, the higher middle schools established in 1886 were remodeled into the *kōtō-gakkō,* or higher schools, in 1894; and in the 20th century these higher schools developed as preparatory schools for the universities.

Higher education was advanced in another area by the College Order of 1903, which enabled certain upper-level private schools to be approved as *semmongakkō,* or colleges, and to receive the same treatment as state-run universities. Until then the private colleges had not been given a clear legal status and had been treated as rather inferior.

Education to 1940. The events of World War I and its aftermath tremendously influenced Japanese society. In the postwar days, Japan experienced the panic and social confusion that was sweeping many nations of the world. Moreover, the intensified leftist movement and the terrible Kantō earthquake of 1923 caused uncertainty and confusion among the Japanese. Nevertheless, the period was one that earned the name of the "Taishō democracy" era, which featured the dissemination of democratic and liberal ideas. It was also a period that marked Japan's real advancement on the world scene and the expansion of its capitalistic economy, all conducive to the flourishing of nationalism. It was quite natural that these social and economic changes should greatly influence education.

Efforts at reform

The Special Council for Education, established in 1917, was charged with making recommendations for school reforms that would adapt the nationalistic education system to the rapid economic growth. Their recommendations involved modifying the existing educational organizations rather than creating new ones. The reform emphasized higher education, though secondary education also grew remarkably. As for elementary education, the target of the reform was to improve the content and methods of education and to establish the financial foundation of compulsory education.

After World War I, the new educational movements generally called progressive in the West were introduced into Japan and came to thrive there. Many private schools advocating this "new education" were established, and the curricula of many state and public schools were also refashioned. The method of new education was gradually introduced into the state textbooks. Preschool education was also encouraged; a state-run kindergarten attached to Tokyo Girls' Normal School had been first established in 1876, and later many public and private kindergartens emerged, particularly after issuance of the Kindergarten Order in 1926.

Government aid for compulsory education was gradually put forward, and by 1940 this developed into a system whereby the government financed half the teachers' salaries and the prefectural governments the other half. Elementary education thus further expanded. Between 1910 and 1940 the number of elementary teachers and pupils almost doubled. In the latter year there were 287,000 teachers and 12,335,000 pupils.

Secondary education continued to be provided by the middle schools for boys, the girls' high schools, and the vocational schools. These schools increased remarkably both in numbers of institutions and in enrollments after World War I, reflecting the social demand. As a result, the secondary schools assumed more of a popular and less of an elitist character than they had evidenced in the Meiji era. In 1931 two courses were provided for the middle-school system; one was for those who advanced on to higher schools, and the other course was for those who went directly on to a vocation. Enrollments of all kinds leaped: whereas in 1910 the enrollments in middle schools, girls' high schools, and vocational schools had been 122,000 pupils, 56,200 pupils, and 64,700 pupils, respectively, the respective figures in 1940 were 432,000 pupils, 555,000 pupils, and 625,000 pupils.

Increases in enrollments, 1910–40

A drastic reform of higher education was instituted in 1918, when the University Order and the Higher School Order were issued on the recommendation of the Special Council for Education. Before that, there had been only the imperial universities, which were state-run. The order approved the founding of private universities and colleges. As a consequence, the old influential private col-

leges, or *semmongakkō,* rich in tradition, were approved as formal universities or colleges, resulting eventually in such famous universities as Keiō and Waseda. National colleges of commerce, manufacturing, medicine, and so on were also opened. In general, universities and colleges multiplied, numbering in 1930 as many as 46 (17 state, five public, and 24 private). College-preparatory education concurrently enlarged through the establishment of public and private higher schools under the Higher School Order. The higher schools were remodeled after the German *Gymnasium* and the French lycée and offered a seven-year course.

The schools could not keep pace with the mounting demand for education. The ratio of applicants to the total number of seats being offered at higher schools, for example, rose from 4.3 in 1910 to 6.9 in 1920 and 10.5 in 1926. Because pupils could not proceed from elementary to secondary schools, and from there to colleges or universities, unless they passed a competitive entrance examination at each stage, the importance and severity of the examinations grew with the number of applicants. Despite efforts by the Ministry of Education to revise and deemphasize the examination system, which was established in the Meiji era, its importance continues to the present day.

After World War I, social education, or education offered outside the formal school system, gained greater recognition in Japan. During the Meiji era, social education, then called "popular education," had been promoted by the Ministry of Education to encourage school enrollment, but by 1890 it had taken the form of adult education, attempting to enlighten middle- and working-class adults with public lectures and library resources. By 1929 social education had again become important as a result of the Ministry of Education's emphasis on youth organizations, supplementary vocational education, youth training, and adult education. The *jitsugyō hoshūgakkō,* or supplementary vocational schools, which had been built after 1893 as part-time educational institutions for working students, reached enrollments exceeding 1,277,000 by 1930. In 1935 *seinengakkō,* or youth schools, were newly established, uniting these supplementary vocational schools with the *seinen kunrenjo,* or youth-training centres, that had earlier been set up to provide military training for youth.

Education changes during World War II. The Manchurian Incident in 1931 escalated into the Sino-Japanese War of 1937, and national life became more and more militaristic. Education acquired an intensely nationalistic character. With the outbreak of war in the Pacific in 1941, the education system underwent emergency "reforms." Elementary schools were renamed *kokumingakkō,* or national schools, under the National School Order issued in 1941. The order proclaimed the idea of a national polity or spirit peculiar to Japan; the content and the methods of education were revised to reflect this nationalism. Moreover, the period of compulsory education was officially extended to eight years, though it actually remained six years because of the worsening war situation.

Militarism and nationalism

Secondary education was similarly made "national." In 1943 the Secondary School Order was issued in an attempt to unify all the secondary schools, but it also, because of the war, shortened secondary education to four years. In the same year the normal school was upgraded to the level of the professional schools. As the war worsened, students above the secondary schools were mobilized as temporary workers in military industries and agricultural communities in order to increase production, and a great number of students were sent to the battlefields. As a result, classes were virtually closed at schools higher than secondary toward the end of World War II.

Education after World War II. On Aug. 14, 1945, Japan accepted the Potsdam Declaration and surrendered unconditionally to the Allied powers. The overriding concern at the general headquarters (GHQ) of the Allied powers was the immediate abolition of militaristic education and ultranationalistic ideology. This was the theme of a directive issued by GHQ to the Japanese government in October 1945. In early 1946 GHQ invited the United States Education Mission to Japan, and it played a decisive role in creating a new educational system. The mission's report recommended thorough and drastic reforms of education in Japan. The report was subsequently adopted in its entirety as the basic framework for a new democratic educational system. The Education Reform Committee, which was directly responsible to the prime minister, was established to make recommendations for the implementation of the new education. Based on these recommendations the Japanese Diet passed a series of legislative acts that forged the foundation of postwar education.

The Fundamental Law of Education and the School Education Law, both enacted in 1947, and the Boards of Education Law of 1948 set the outlines of the new education. The prewar system was replaced by a democratic single-track system, in which school programs were integrated and simplified and the period of attendance was settled in six, three, three, and four years, respectively, for *shōgakkō,* or elementary schools, *chūgakkō,* or lower secondary schools, *kōtōgakkō,* or upper secondary schools, and *daigaku,* or universities. The period of compulsory attendance was extended to nine years; coeducation was introduced; and provisions were made for education for the physically handicapped and other special education.

The new Japanese school system

The reform of the content of education proceeded to reduce the strong state control of former days and to encourage teachers' initiative. State textbooks were abolished in favour of commercial ones, and schools were controlled locally by elective boards of education. *Shūshin* disappeared from the curricula and was replaced by new subjects, such as *shakaika,* or social studies, designed to prepare children for life in a democratic society. The educational reform also altered the character of the universities, which offered access to all citizens. The former institutions—universities, colleges, and normal schools—were reorganized into four-year universities and colleges. Teacher education was placed within the university system, and anyone who completed professional training was eligible for teacher certification. This reorganization had an immense impact upon the development of higher education.

The peace treaty of 1952 not only liberated Japan from the restraints of occupation but also allowed education there to be adjusted to intrinsic cultural and political orientations. Centralization of control increased with respect to administration, curriculum, textbooks, and teacher performance through a series of legislative and administrative measures in the 1950s. In addition, the political indoctrination of the leftist Japan Teachers' Union was hindered, and moral education was reintroduced as a requirement at the elementary and lower secondary levels. On the whole, however, the postwar educational reforms were retained and advanced, and their subsequent elaboration helped match Japan's rapid economic growth.

Changes after the end of Allied occupation

The postwar educational administration was organized into a three-tiered structure, with national, prefectural, and municipal components—all under the general supervision of the Ministry of Education, which also wields a considerable measure of authority over curricular standards, textbooks, and school finance, among other functions. Through its central, advisory role, the Ministry of Education has guided the development of egalitarian and efficient schooling in the postwar era.

The progressive curriculum, which emphasized child interest and was introduced from the United States immediately after the war, produced deteriorating student performance. Thus, during 1961–63 the Ministry of Education replaced that curriculum with a discipline-centred curriculum at the elementary and lower secondary levels in order to improve academic achievement, moral education, science and technical education, and vocational education. This curricular revision set the tone for later changes in the national curriculum. Each major curricular revision represented an educational response to a variety of social needs, above all economic.

The 1960s was a period of high growth for both the economy and education. The unprecedented economic growth was stimulated by an ambitious national plan to boost individual income, industry, and trade. Responding to the changing economic and industrial environment, enrollments in high schools and in colleges or universities increased, respectively, from 57.7 and 10.3 percent of

the eligible students in 1960 to 91.9 and 37.8 percent in 1975. Ninety percent of this increase in university and college enrollments was absorbed into poorly financed private institutions, which contributed to the deterioration of higher education. Problems also arose at the upper secondary level, where education remained rigidly uniform, even though students were increasingly diverse in ability, aptitudes, and interests. The inability of the postwar educational system to meet either student requirements or the insatiable demands for secondary and postsecondary education became of critical concern, and in 1971 the Central Council for Education recommended reforming Japan's education to eradicate these problems.

The Central Council initiated a sustained school reform debate that set the stage for the establishment, in 1984, of an advisory council on educational reform, which is directly responsible to the prime minister. The advisory council called for elimination of the uniformity and rigidity of education at all levels and for the enhancement of "individuality" through education. Its recommendations in 1987 included diversifying upper secondary education, improving moral education, encouraging greater local freedom and responsibility in developing curriculum, improving teacher training, and fostering diversity in higher education. Thus, Japan's educational policy is being directed toward meeting the diversified needs of the future. (A.Na./N.S.)

South Asia. *Preindependence period.* Amid the rising nationalism of the latter part of the 19th century, Indians became more and more critical of the domination of Western learning as imposed by the British rulers and demanded instead more attention to Indian languages and culture. The Indian National Congress, several Muslim associations, and other groups raised their voices against the British system of education. Nor were British authorities altogether blind to the needs of the country. When Baron Curzon of Kedleston arrived as viceroy in 1898, his determination to improve education was immediately translated into an order for a close survey of the entire field of education. It revealed: "Four out of five villages are without a school. Three boys out of four grow up without any education and only one girl out of forty attends any kind of school." Education had advanced, but it had not penetrated the country as the British had earlier expected.

Political conflicts involved in education in India

Curzon applied himself to the task of putting matters in order. He disapproved of the doctrine of state withdrawal and instead considered it necessary for the government to maintain a few institutions of every type as models for private enterprise to imitate. He also abandoned the existing policy of educational laissez-faire and introduced a stricter control over private schools through a vigilant policy of inspection and control. Such a policy aroused bitter feelings among some educated Indians, since it was believed that Curzon was bent on bringing the entire system of education under government control.

The main battle, however, was fought over the universities. With Eton and Balliol in mind, Baron Curzon set up the Indian Universities Commission of 1902 to bring about a better order in higher education. The commission made a number of important recommendations—namely, to limit the size of the university senates; to entrust teaching in addition to examining powers to universities; to insist on a high educational standard from affiliated colleges; to grant additional state aids to universities; to improve courses of studies; to abolish second-grade colleges; and to fix a minimum rate of fees in the affiliated colleges. The report was severely criticized, and the last two recommendations had to be dropped. Legislation in regard to the other proposals was passed despite bitter opposition in the legislature and the press.

The conflict resulted less from educational differences than from political opinions on centralization. In one part of the country, violent agitation had already started on the question of the partition of Bengal. In another, the patriot Bal Gangadhar Tilak declared: "Swaraj [self-rule] is our birthright." Thus, Baron Curzon's educational reforms were considered sinister in their intentions, and his alleged bureaucratic attitude was resented.

The administrative policy of Baron Curzon also gave rise to the first organized movement for national education. This effort was part of the swadeshi movement, which called for national independence and the boycotting of foreign goods. A body known as the National Council of Education, in Calcutta, established a national college and a technical institution (the present Jadavpur University) in Calcutta and 51 national schools in Bengal. These schools sought to teach a trade in addition to ordinary subjects of the matriculation syllabus. The movement received a great impetus, because the Calcutta Congress (1906) resolved that the time had arrived for organizing a national system of education. With the slackening of the swadeshi movement, however, most of the national schools were eventually closed. The effect of the movement was nevertheless noticeable elsewhere: Rabindranath Tagore started his famous school in West Bengal near Bolpur in 1901; the Arya Pratinidhi Sabha established *gurukulas* at Vrindāban and Hardwār; the Indian National Congress and the All-India Muslim League at their sessions in Allahābād and Nāgpur, respectively, passed resolutions in favour of free and compulsory primary education.

Swadeshi movement

In 1905 Baron Curzon left India. In order to pacify the general public, his successors modified his policy to some extent, but the main program was resolutely enforced. Although Indian public opinion continued its opposition, the reforms of Baron Curzon brought order into education. Universities were reconstituted and organized, and they undertook teaching instead of merely conducting examinations for degrees. Colleges were no longer left to their own devices but were regularly visited by inspectors appointed by the universities. The government also became vigilant and introduced a better system for inspecting and granting recognition to private schools; the slipshod system of elementary education was also improved. The number of colleges and secondary schools continued to increase as the demand for higher education developed.

In 1917 the government appointed the Sadler Commission to inquire into the "conditions and prospects of the University of Calcutta," an inquiry that was in reality nationwide in scope. Covering a wide field, the commission recommended the formation of a board with full powers to control secondary and intermediate education, the institution of intermediate colleges with two-year courses, the provision of a three-year degree course after the intermediate stage, the institution of teaching and unitary universities, the organization of postgraduate studies and honours courses, and a greater emphasis on the study of sciences, on tutorial systems, and on research work. The government of India issued a resolution in January 1920, summarizing the report of the commission. Since then, all legislation of any importance on higher education in any part of India has embodied some of the recommendations of the commission.

Meanwhile, World War I had ended, and the new Indian constitution in 1921 made education a "transferred" subject (that is, transferred from British to Indian control), entrusting it almost entirely to the care of the provinces. In each province, educational policy and administration passed into the hands of a minister of education, responsible to the provincial legislature and ultimately to the people. Although European-style education was still maintained as a "reserved" subject and was not placed under the control of the Indian minister of education, this anomaly was corrected by the Government of India Act of 1935, which removed the distinction between transferred and reserved subjects and introduced a complete provincial autonomy over education.

Generally, the new constitution of 1921 was considered inadequate by the Indian National Congress. In protest, Mahatma Gandhi launched the Non-cooperation Movement, the campaign to boycott English institutions and products. National schools were established throughout the country, and *vidyapeeths* ("national universities") were set up at selected centres. The courses of study in these institutions did not differ much from those in recognized schools, but Hindi was studied as an all-India language in place of English, and the mother tongue was used as the medium of instruction. These institutions functioned for a short time only and disappeared with the suppression of

Non-cooperation Movement

the Non-cooperation Movement. The Congress' struggle for self-rule, however, became more vigorous, and with it spread the national movement toward education to suit national needs. The Government of India Act of 1935 further strengthened the position of the provincial ministers of education, since the Congress was in power in major provinces. The developmental program of provincial governments included the spread of primary education, the introduction of adult education, a stress on vocational education, and an emphasis on the education of girls and underprivileged people. The importance of English was reduced, and Indian languages, both as subjects of study and as media of instruction, began to receive greater attention.

General educational trends, 1921–47

On this general background, educational developments from the inauguration of reforms in 1921 until independence in 1947 can be viewed. In the field of elementary education, the most important event was the passing of compulsory-education acts by provincial governments—acts empowering local authorities to make primary education free and compulsory in the areas under their jurisdiction. Another noteworthy feature was the introduction of Gandhi's "basic education," which was designed to rescue education from its bookish and almost purely verbal content by emphasizing the teaching of all school subjects in correlation with some manual productive craft. A general demand for secondary education developed with the political awakening among the masses. Schools in rural, semi-urban, and less advanced communities were established, as well as schools for girls. Some provision was made for alternative or vocational courses when the provincial governments started technical, commercial, and agricultural high schools and gave larger grants to private schools providing nonliterary courses. But the expected results were not achieved because of a lack of funds and of trained teachers. Secondary schools still concentrated on preparing students for admission to colleges of arts and sciences.

The period is also marked by a diminishing of the prejudices against the education of girls. The impetus came from the national movement launched by Gandhi, which led thousands of women to come out of the purdah for the cause of national emancipation. It was also realized that the education of the girl was the education of the mother and through her of her children. Between 1921–22 and 1946–47, the number of educational institutions for girls was nearly doubled.

In the field of university education, outstanding developments included (1) the establishment of 14 new universities, unitary as well as affiliating, (2) the democratization of the administrative bodies of older universities by a substantial increase in the number of elected members, (3) the expansion of academic activities through the opening of several new faculties, courses of studies, and research, (4) a substantial increase in the number of colleges and student enrollments, (5) the provision of military training and greater attention to physical education and recreational activities of students, and (6) the constitution of the Inter-University Board and the development of intercollegiate and interuniversity activities. With these improvements, however, the educational system of the country had become top-heavy.

The postindependence period in India. India and Pakistan were partitioned and given independence in 1947. Since then, there has been remarkable improvement in scientific and technological education and research, but illiteracy remains high (less than 40 percent of Indians aged four and older are literate). The new constitution adopted by India did not change the overall administrative policy of the country. Education continues to be the prime responsibility of the state governments, and the union (central) government continues to assume responsibility for the coordination of educational facilities and the maintenance of appropriate standards in higher education and research and in scientific and technical education.

In 1950 the government of India appointed the Planning Commission to prepare a blueprint for the development of different aspects of life, education being one of them. Since then, successive plans (usually on a five-year basis) have been drawn and implemented. The main goals of these plans have been to achieve universal elementary education; to eradicate illiteracy; to establish vocational and skill training programs; to upgrade standards and modernize all stages of education, with special emphasis on technical education, science, and environmental education, on morality, and on the relationship between school and work; and to provide facilities for high-quality education in every district of the country.

The education commissions and their recommendations

Since 1947 the government of India has also appointed three important commissions for suggesting educational reforms. The University Education Commission of 1949 made valuable recommendations regarding the reorganization of courses, techniques of evaluation, media of instruction, student services, and the recruitment of teachers. The Secondary Education Commission of 1952–53 focused mainly on secondary and teacher education. The Education Commission of 1964–66 made a comprehensive review of the entire field of education. It developed a national pattern for all stages of education. The commission's report led to a resolution on a national policy for education, formally issued by the government of India in July 1968. This policy was revised in 1986. The new policy emphasizes educational technology, ethics, and national integration. A core curriculum was introduced to provide a common scheme of studies throughout the country.

The national department of education is a part of the Ministry of Human Resource Development, headed by a cabinet minister. A Central Advisory Board of Education counsels the national and state governments. There are several autonomous organizations attached to the Department of Education. The most important bodies are the All-India Council of Technical Education (1945), the University Grants Commission (1953), and the National Council of Educational Research and Training (1961). The first body advises the government on technical education and maintains standards for the development of technical education. The second body promotes and coordinates university education and determines and maintains standards of teaching, examination, and research in the universities. It has the authority to enquire into the financial methods of the universities and to allocate grants. The third body works to upgrade the quality of school education and assists and advises the Ministry of Human Resource Development in the implementation of its policies and major programs in the field of education.

The central government runs and maintains about 1,000 central schools for children of central government employees. It has also developed schools offering quality education to qualified high achievers, irrespective of ability to pay or socioeconomic background. The seventh five-year plan (1985–90) specified that one such *vidyalaya* would be set up in each district. The state governments are responsible for all other elementary and secondary education. Conditions, in general, are not satisfactory, although they vary from state to state. Higher education is provided in universities and colleges.

From the 1950s to the '80s the number of educational institutions in India tripled. The primary schools, especially, experienced rapid growth because the states have given highest priority to the universalization of elementary education in order to fulfill the constitutional directive of providing universal, free, and compulsory education for all children up to the age of 14. Most but not all children have a primary school within one kilometre of their homes. A large percentage of these schools, however, are understaffed and do not have adequate facilities. The government, when it revised the national policy for education in 1986, resolved that all children who attained the age of 19 years by 1990 would have five years of formal schooling or its equivalent. Plans have also been made to improve or expand adult and nonformal systems of education. Dissension among political parties, industrialists, businessmen, teacher politicians, student politicians, and other groups and the consequent politicization of education have hampered progress at every stage, however.

The postindependence period in Pakistan. On Aug. 14, 1947, Pakistan emerged as a national sovereign state. For the new state the initial years proved to be a period essentially of consolidation and exploration. The constitution adopted in 1956 recognized the obligation of the state to

provide education as one of the basic necessities of life. The new constitution implemented by the National Assembly in 1973 made practically no changes to the original educational policy. The federal Ministry of Education, headed by the federal education secretary, oversees education in the federal capital territory and in national institutions and determines policies and standards. Provincial governments handle all other administrative duties.

Aims of Pakistani educational policy

Beginning in 1955, Pakistan adopted a series of five-year plans to improve economic and educational development. The most important educational objectives of the sixth plan (1983–88) were: (1) to strengthen training programs for all categories of manpower, (2) to establish technical trade schools and vocational institutes, (3) to provide adequate machinery, materials, and books for workshops, laboratories, and other facilities, and (4) to strengthen and develop centres for advanced engineering studies. Because less than 30 percent of the adult population is able to read and write, literacy is also a major area of concern. The National Education Policy of 1979 emphasized the need for improving vocational and technical education and for disseminating a common culture based on Islāmic ideology. It also announced plans for gradually replacing the four-tier school structure (primary, secondary, college, and university) with a three-tier system consisting of primary (grades one through eight), secondary (grades nine through 12), and higher education.

The government has accepted responsibility for providing free primary education for a length of time fixed provisionally at five years. Only a little more than 50 percent of primary-age children are enrolled in schools, however, with attendance concentrated in urban areas. Religious classes providing Islāmic moral and sociocultural education have been taught in the schools since about 1980. An alternative course for non-Muslim students is also being introduced.

The postindependence period in Bangladesh. Comprising what was formerly the eastern wing of Pakistan, Bangladesh emerged as an independent sovereign state in December 1971. Thus, it shares its educational history with India until 1947 and with Pakistan from 1947 to 1971. After independence Bangladesh continued to follow the primary education scheme originally established by Pakistan. One of the country's most valued educational assets is its rich national language, Bengali.

Article 17 of the constitution of the People's Republic of Bangladesh declares that it is the duty of the state to provide education to all its children to such stage as may be determined by law. In 1973 and 1974 the government nationalized most of the primary schools, but it was found that about 33 percent of primary-school-age children in Bangladesh never went to school and that about 70 percent of those who did left school before attaining the minimum educational standard. The majority of children thus enter adulthood illiterate. It has now been recognized that universalization of primary education for an overpopulated developing country like Bangladesh is a difficult task. Major reforms are under way to orient the educational system to a new social order inspired by the ideals of "nationalism, democracy, socialism, and secularism" on which the new nation is founded.

The postindependence period in Sri Lanka. Sri Lanka (formerly Ceylon) gained independence in 1947. Successive governments have since continued the policy of democratizing education that began under British rule. The political and social changes ushered in during the preindependence period paved the way for a gradual process of constitutional reforms. Schools and schooling are seen as great instruments of socioeconomic development.

Education is free from the kindergarten to the university level in all state and state-aided institutions. Although there are a few fee-levying private institutions, management of education is primarily a state responsibility. General education within the formal system is divisible into primary, junior secondary, and senior secondary education. There are few dropouts or grade repeaters at the primary level. Thus, the percentage of literacy rose from 57.8 (70.1 for males and 43.8 for females) in 1948 to 86.5 (90.5 males, 82.4 females) in 1981. At the junior secondary stage, instruction is provided according to a common curriculum that consists of religion and other subjects. Students at the senior secondary stage are streamed into science, commerce, or liberal arts courses.

The University Act of 1978 established the University Grants Commission and the University Services Appeals Board to provide for the establishment, maintenance, and administration of universities and other higher educational institutions together with their campuses and faculties. The National Institute of Education was established in 1987 to coordinate curriculum development, textbook development, teacher education, and eventually certification and entrance examinations. (S.N.M.)

Africa. Before the arrival of the European colonial powers, education in Africa was designed to prepare children for responsibility in the home, the village, and the tribe. It provided religious and vocational education as well as full initiation into the society. In sub-Saharan Africa it varied from the simple instruction given by fathers to children among the San of the Kalahari to the complex educational system of the sophisticated and highly organized Poro society of western Africa (extending over Liberia, Sierra Leone, and Guinea). The majority of ethnic groups in Africa fell somewhere between the San and the Poro with respect to the educational arrangements they provided for their youth. Most societies offered rituals to mark the end of puberty and relied heavily upon custom and example as the principal educational agents. The rites of passage marked the culmination of an epoch in a boy's life. As a child, he had been introduced by his elders to the legends surrounding previous exploits of his tribe, to the mysteries of his religion, to the practical aspects of hunting, fishing, farming, or cattle-raising, and to his community responsibility. Now he occupied a new position in the society. In some cases he had been prepared for the rites; in others secrecy surrounded the event, for reaction to the ceremony was itself an important part of the ritual. A variety of formal observances, in addition to the experiences of daily living, impressed upon the youth his place in the society, a society in which religion, politics, economics, and social relationships were inextricably interwoven. Girls underwent a similar, though usually shorter, initiation period.

Traditional African education

An exception to this pattern could be found in those areas where Islām had spread. Islām reached eastern Africa in the 9th and 10th centuries and western Africa in the 11th. It introduced the Arabic script, and, because knowledge of the Qur'ān became an important religious requirement, Qur'ānic schools developed. These schools concentrated on the teaching and memorization of the Qur'ān; some were little more than gathering places beneath a tree where teachers held classes. Qur'ānic schools placed young Africans in contact with Arab civilizations, and boys selected as potential leaders could attend higher educational institutions in the Arab world. Nevertheless, Islām touched but a small fraction of the total African population of sub-Saharan Africa.

Western-style schooling was introduced in most of Africa after the establishment of the European colonial powers. As African nations gained independence in the late 20th century, they abolished the racial segregation that had existed and instituted other reforms but, in general, kept the structure of the existing school systems, at least initially. Thus, 20th-century education in these countries can be discussed according to former colonial status. Education in Ethiopia, Liberia, and South Africa, however, must be treated separately—Ethiopia and Liberia because they have long histories as independent nations and South Africa because it remains under the control of a white minority government.

Ethiopia. Christianity was recognized in Ethiopia in the 4th century. For nearly 1,500 years all education was church-related and hence church-controlled, except in the eastern part of the country where the Islāmic population maintained Qur'ānic schools. In 1908 Emperor Menelik II created the embryonic government school system, modeling it on European systems. The real development of education, however, came after World War II under the direction of Emperor Haile Selassie. Despite his efforts, by 1969 less than 10 percent of the children between the

Post-World War II expansion of Ethiopian education

ages of seven and 12 were in school. Education at the secondary level benefited from the infusion of more than 400 Peace Corps teachers in the 1960s and early 1970s. In the 1950s the first Ethiopian colleges were founded. By 1970, 2,800 Ethiopian students were enrolled in higher education either in their own country or overseas.

In 1974 a military revolution overthrew the emperor. Ethiopia declared itself a socialist state and proclaimed that socialism would permeate all aspects of the society. The government's stated aims of education were (1) education for production, (2) education for scientific consciousness, and (3) education for social consciousness. Political alliance with the Soviet Union influenced educational reform. Polytechnical education, which emphasizes familiarizing children with the important branches of production and acquainting them with first-hand practical experience, was widely introduced by Soviet educational advisers. A number of Ethiopian students have been sent to the Soviet Union or Eastern-bloc countries for higher education or to Cuba for schooling at the secondary level.

The structure of the Ethiopian school system remained unchanged from that established in the late 1950s. Children begin the 12-year program at age seven. Grades one through six make up the primary cycle, seven through eight the junior secondary cycle, and nine through 12 the senior secondary cycle. Students who pass the Ethiopian School Leaving Certificate Examination at the end of grade 12 are eligible for higher education, but space in the country's colleges and universities is limited.

Liberia. Education in Liberia, the oldest republic in Africa (1847), is distinctly different from that in any other African country. Liberia was founded by freed slaves from the United States, and its educational system was modeled after the American system. Public primary and secondary schools were established in the 19th century for the children of the settlers, but there was little money to extend schooling into the interior of the country for the indigenous people. Church schools were also established. The Western-style schools trained Liberians in the new settlements for work in offices. A few students were prepared for the legal or theological profession.

American influence on Liberian education

In 1912 a centralized educational system was established under a Cabinet-level official, but, except for the establishment of a few secondary schools and colleges, nothing of importance happened until the end of World War II. In the prewar period three-fourths of the schools were either private or mission-run. Economic growth and the interest of President William V.S. Tubman in the 1950s resulted in a greater extension of education for indigenous Liberians. The educational system was organized to provide preprimary education for children aged four and five years, six years of elementary education for children aged six to 12, and three years each of junior and senior high school. Postsecondary education can be pursued at three leading institutions: the University of Liberia, sponsored by the government; Cuttington University College, administered and financially supported by the Episcopal church with some financial aid from the government; and the William V.S. Tubman College of Technology. The educational expansion started by President Tubman in the 1950s has, however, due to the lack of finances, reached only a small fraction of the people. (Da.G.S.)

South Africa. From the time of the first white settlements in South Africa, the Protestant emphasis on home Bible reading ensured that basic literacy would be achieved in the family. Throughout the development from itinerant teachers to schools and school systems, the family foundation of Christian education remained, though it was gradually extended to embrace an ethnic-linguistic "family."

Early efforts to establish education in South Africa

Despite some major 19th-century legislation on the administration of education (1874 in the Transvaal and the Orange Free State, 1865 in Cape of Good Hope, 1873–77 in Natal) and some early efforts to establish free schools, political and linguistic problems impeded the development of public education before 1900. Natal had gone furthest in affirming government responsibility for education and setting up the necessary administrative machinery, but, by and large, provision for schooling remained voluntary and piecemeal until the beginning of the 20th century.

The Boer, or South African, War (1899–1902) suspended educational development entirely and confirmed the resolve of each white South African group to protect its own cultural prerogatives. When the Union of South Africa was created in 1910, it was a bilingual state, and, in education, English-speaking and Afrikaans-speaking schools were established for white Europeans. Furthermore, a political tightness and separateness increased among the Afrikaners after the war and strengthened their tendency to exclude nonwhites from the cultural and political life of the dominant society. The trend toward separate schools for linguistic and racial groups became a rigid practice in most of South Africa after union.

Church mission schools attempted to replace the preliterate tribal education of native Africans in the South African colonies. Established from 1789, they were dedicated to converting the natives to Christianity and generally inculcating an attitude of service and subservience to whites. These schools spread from 1823 to 1842, and colonial governments made occasional grants to them from 1854. Some mission schools included a mixture of races, but, by and large, segregation was established by custom. Although some exemplary schools followed rather liberal social and curricular policies, most schools held to narrowly religious content. The mission schools were virtually brought into the state system through government subsidies and through provincial supervision, inspection, and control of teaching, curriculum, and examination standards.

The union and school systems

By the time the union was formed, the new provinces had each established school systems, structured mainly for European children but including provisions for other groups. Specific arrangements varied, but basically the systems were headed by a department of education under a director and controlled through an inspectorate. Three of the provinces had school boards that localized the department administration. Compulsory-attendance regulations were being effected for European children, while separate school developments were under way for other groups. The language of instruction had been established provincially, with both Afrikaans and English in use.

The South Africa Act of 1909 left the control of primary and secondary education with the provinces, while reserving higher education to the union government. The Union Department of Education, Arts, and Science became the central educational authority and expanded its responsibilities by accepting control of special sectors such as vocational, technical, and artistic education. In 1935 an Interprovincial Consultative Committee was established to coordinate educational matters among the provinces and between the provinces and the central government. In 1967 the union government passed the National Education Policy Act, which, with the Amendment Act of 1982 and the Constitution Act of 1983, forms the basis for South African education. The provinces have incorporated this policy into their own legislation and administrative provision.

Administration of education is divided between national departments and provincial authorities. Because education is differentiated by race, four separate systems must be distinguished, although recent legislation has stated the principle of equal education opportunity and structures have become similar for white, Coloured (of mixed ancestry), and Indian (Asian) population groups. Education for whites is under control of the Minister of National Education, and provincial federal coordination is accomplished through a National Education Council and a Committee of Heads of Education. Education for Coloureds and Indians is administered through the legislative bodies representing these groups, the House of Representatives and the House of Delegates, respectively. Education for blacks is the responsibility of the four independent (although not internationally recognized) republics (Transkei, Bophuthatswana, Venda, Ciskei) and the six nonindependent "states" (Gazankulu, KaNgwane, KwaNdebele, KwaZulu, Lebowa, and Qwaqwa) established by the government. For blacks not in these areas, the Department of Education and Training administers education.

All four systems are now supposed to be following the same basic organizational and curricular patterns. Legis-

lation in 1973 and 1983–84 has made it possible to refer to the education of Coloured and Indian children when describing structural features of schooling developed for white children. The system is organized into four three-year cycles: junior primary, senior primary, junior secondary, and senior secondary. Because the first year of the junior secondary cycle is taken in the primary school, the primary and secondary units are seven and five years respectively (replacing the earlier eight–four organization). Schooling is compulsory from age seven to 16. Core syllabi are established by the National Education Council and elaborated by the provinces or other departments. The general high schools are predominantly academic but offer a range of streams. Specialized high schools, at the senior secondary level, offer technical, agricultural, commercial, art and domestic science courses. Apprenticeship may be begun after the first year of the senior secondary phase (grade 10). Attempts are now being made to form regional comprehensive schools. Private schools, both independent and aided, are found mainly in Transvaal and Cape Province, but over 90 percent of South African white children are in state schools.

English and Afrikaans and the division of education

The language of instruction does not greatly affect the curriculum or organization of schools, but it does present a basic educational division in South Africa. Both English and Afrikaans are official languages of the republic. About 60 percent of the white population speak Afrikaans and about 40 percent English, and the two languages are used almost proportionately in schools. By law, instruction is given in the mother tongue, the other language being introduced as a second language. Coloured and Asiatic pupils are taught in the official language common to their area of residence, normally Afrikaans and English respectively. African pupils are taught in their mother tongue at least until they begin the senior primary cycle (grade 4), and then one of the official languages is used. They must study both English and Afrikaans, although they may offer one African and one official language for the examination at the end of the secondary course. Language is intimately related to politics and to African aspirations. It was the imposition of Afrikaans as the compulsory language of instruction that triggered the Soweto riots in 1976 and the subsequent wave of unrest. Black parents and students want recognition of their own language and culture (Africanization) as well as the access to the metropolitan culture of their own and other countries that English gives.

In 1922, when the Phelps–Stokes Commission on education in Africa offered its report, South Africa's example in the development of liberal and adaptable educational provisions for Africans, particularly in Natal and Cape Province, was held up for emulation. The passing of the tribal system was noted and efforts toward interracial cooperation complimented. It was obvious, however, that little of value to Africans was being done in the European-model schools and that noteworthy educational efforts were associated with special institutions, such as Lovedale School and University College of Fort Hare in the Cape.

Concern over Bantu education in the 1940s led to the creation of the Eiselen Commission, whose report in 1951 accorded with the separatist views of the nationalist government coming to power in 1948 and presented a basis for apartheid legislation in education. The Bantu Education Act of 1953 resulted.

Apartheid and Bantu education

The policy of apartheid, or "separate development," has been controversial, both within South Africa and throughout the world. Fundamentally, the government position rests on three assumptions: (1) that each cultural group should be encouraged to retain its identity and develop from its present stage according to its unique characteristics, (2) that, with a population of diverse racial-social groups, the way to ensure peaceful coexistence and general progress is through legal and institutional separation, and (3) that the only agency capable of exercising overall responsibility for this development is the central government. Implementation of apartheid policy has led to a virtually total separation of educational facilities for white, black, Coloured, and Indian populations, with resulting divergence of opportunity between the extremes of black and white education.

The Education and Training Act of 1979 established principles following the pattern of white education for the education of blacks. Slightly different school organization, designation of state-aided community schools with school committees, provision for African language instruction, and separate administration are formal characteristics distinguishing the system for blacks. More important are the effects of inequality on the system's operation. Although the government has introduced a limited experiment in compulsory education, the dropout rate among blacks is high from the earliest years, and markedly so after the fourth year. Because of work conditions, pupils may be educated in factory, mine, or farm schools that are less adequate than general schools. Teacher qualifications are lower for blacks than for the other groups. Illiteracy is high. Rural schools are crowded and short of materials. Although black pupils constitute about 70 percent of the primary school enrollment, few attend secondary schools.

There have been some attempts made to close the gap between black and white education at both lower and higher levels. The government has proclaimed the principle of equal educational opportunity and has increased financing, private and community efforts have augmented schooling and introduced experimental integrated schools, and private schools and white universities have opened to black students. Black schools remain severely inadequate, however; and the government's position that the immensity of the problem defies immediate solution conflicts with the demands of black activist student organizations, which have multiplied since 1976 (partly through division) and intensified their resistance through strikes and boycotts. Thus, violence and fear have increasingly intruded on township schools and on black universities.

South African university education

The tertiary sector of South African education includes universities, technikons—successors to the colleges of advanced technical education, offering programs of one to six years in engineering and other technologies, management, and art—technical colleges and institutes, and colleges of education. Technical centres, industrial training centres, and adult education centres extend training to early school-leavers. The previously discriminatory qualifications required for primary and secondary teachers as well as for teachers from the different racial groups have been standardized. All teachers must complete the full secondary course plus a three-year training course. Teachers' professional organizations comprise a bewildering array of English, Afrikaaner, African. Coloured, and Indian organizations on the national and local levels. Recently, there has been agitation for a single teacher-registering body. Joint bodies already exist.

By the Extension of University Education Act in 1959, nonwhites were barred from entrance to white universities and separate university colleges were set up on an ethnic-linguistic basis. This well-organized system of differentiating groups has begun to break down, however, as first English, then Afrikaans universities stated their policies of admission by merit, as university decisions and legislation have opened nonwhite universities to other groups, and protests against government quotas on university admissions have become increasingly effective. The universities have become centres of agitation against apartheid.

A major government commission, conducted through the Human Sciences Research Council, in 1981 recommended the establishment of a single system of education under a single ministry. Although principles of the report were accepted, the government has held to a cultural policy from which institutional separation is derived. The change from an ideological basis to a pragmatic basis for this separation, combined with the elimination of formal barriers to racial crossovers and black mobility in education, has produced a policy that is competing with revolutionary strategies for social change. (R.F.L.)

Early mission and lay schools

General influences and policies of the colonial powers. During the colonial period, the first direct "educational" influences from outside came from religious missionaries, first Portuguese (from the 15th century) and then French, Dutch, English, and German (from the 15th to the 19th century). Starting from coastal bases, they undertook to penetrate into the interior and begin campaigns to convert

the black populations. The missions were the first to open schools and to develop the disciplined study of African languages, in order to translate sacred texts or to conduct religious instruction in the native tongues.

The partition of Africa by the colonial powers in the 19th and early 20th centuries led first to the establishment of mission schools and then to the establishment of "lay" or "public" or "official" schools. The importance of either the lay or the religious system depended on the political doctrines of the mother country, its institutions (a firmly secular state or one with a state religion), and the status of the colony and its history. But, whatever the system, the fundamental purpose of colonial instruction was the training of indigenous subaltern cadres—clerks, interpreters, teachers, nurses, medical assistants, workers, and so forth—all indispensable to colonial administration, businesses, and other undertakings. Though inspired by the system in the mother country, no colonial system was equivalent to its prototype. The intention was not to "educate" the subject peoples but to extend the language and policies of the colonizer.

Such a generalization, though, is subject to a slight qualification with reference to the religious missions. Both the missions and the political administrations wished to model the African man in accordance with their own needs and objectives. The religious missions, however, became involved in the cultures of the Africans through continual contact with them in the daily ministrations; they used African languages in instruction wherever the colonial administration permitted it. Moreover, for a long while, religious establishments were alone in offering vocational education, some secondary education, and even some higher education to Africans—frequently in the face of the fears or opposition of the colonial authorities.

Education in Portuguese colonies and former colonies. Angola and Mozambique shared a common historical legacy of hundreds of years of Portuguese colonization, and the general overall educational philosophy for both countries was the same until independence. For Portugal, education was an important part of its civilizing mission. In 1921, Decree 77 forbade the use of African languages in the schools. The government believed that since the purpose of education was integration of Africans into Portuguese culture the use of African languages was unnecessary. In 1940 the Missionary Accord signed with the Vatican made Roman Catholic missions the official representatives of the state in educating Africans. By the 1960s an educational pattern similar to that in Portugal had emerged. It began with a preprimary year in which the Portuguese language was stressed, followed by four years of primary school. Secondary education consisted of a two-year cycle followed by a three-year cycle. After 1963 two universities were opened, one in Angola and the other in Mozambique. In addition, postprimary education was offered in agricultural schools, in nursing schools, and in technical service courses provided by government agencies. Despite remarkable progress in the 1960s, primary education was available to few Africans outside urban areas, and even there, the proportion of African children in secondary schools was low.

Educational reforms after independence

Marxist governments triumphed in both Angola and Mozambique when independence came in 1975. Dissident groups, however, have maintained bloody civil wars in both countries that have had disastrous effects on the educational systems. The Popular Liberation Movement of Angola (Movimento Popular de Libertação de Angola; MPLA), which gained control of Angola when Portugal withdrew, had educational reform as one of its main objectives even during the fight for independence. A report of the first congress of the MPLA published in 1977 provided a blueprint that has been followed with few deviations. Marxism-Leninism is stressed as the base for the educational system. The training of all people to contribute to economic development is a major objective. Eight years of primary education is to be universal. Secondary education, offered on a limited basis, includes vocational as well as college preparatory courses. At the University of Angola special emphasis has been placed on scientific and engineering courses.

The governing Mozambique Liberation Front (Frente de Libertação de Moçambique; Frelimo) introduced its educational system in the areas it controlled even before independence. After independence, at the Third Congress of Frelimo in February 1977, policies for the transition to socialism were formalized. While Marxism would provide a foundation, the particular needs of Mozambique would be addressed. All schools were nationalized, but because most of the teachers, who were Portuguese, had left the country, the government was faced with a tremendous teacher shortage. Crash programs in teacher training were introduced. Textbooks, although very limited, were produced that reflected the culture of Mozambique. Most are in Portuguese, which remained the official language of the country, in part because none of the multitude of different cultural groups dominates.

German educational policy in Africa. Well before Chancellor Otto von Bismarck had granted a charter to the German Colonial Society in 1885, German missionaries, both Protestant and Catholic, were operating in various regions of western, central, and eastern Africa—from 1840 in Mombasa (now in Kenya), from 1845 in Cameroon, from 1847 in Togo, and from 1876 in Buganda (now Uganda) and in Mpwapwa and Tanga (now in Tanzania). Instruction was everywhere conducted in the local languages, which were objects of study by numerous missionaries and by eminent scholars.

Missionary schools in German areas

On the eve of World War I, more than 95 percent of the schools in German Africa were operated by religious groups. In the southwestern part of the continent the government did not establish any schools at all, relying completely on missionary activity. (In eastern Africa, however, where the large Muslim population was unwilling to send its children to schools managed by Christian religious groups, the government did assume a more active educational role.) To assist the missions, the government granted aid to those schools that met requirements based on specific government needs that changed with time. An example of this sort of aid was the fund founded in 1908 for the dissemination of the German language. The missions had not previously been required to include German in the curriculum but were now forced to do so in order to receive money from the new fund. The language problem was a persistent one and was handled differently in different colonies. In eastern Africa, Swahili was recognized as a language and emphasized in the lower schools, thus providing a lingua franca for the entire area. The government attempted a similar policy with Ewe in Togo and Douala in the Cameroons, but, in southwestern Africa, German was the language of instruction.

Throughout the literature on German educational policy in the African colonies, there is a continued emphasis on the necessity for vocational education and practical work. The missions, however, were more interested in establishing schools providing general education, and lay German educators took a dualistic approach to African education, emphasizing both practical and academic studies.

The absorption of German colonies by England and France after World War II eradicated most of the German influence in education. However, the German insistence on Swahili in German East Africa left that area far more unified linguistically than any other colonial area.

Education in British colonies and former colonies. In the British colonies, as elsewhere, religious missions were instrumental in introducing European-type education. The Society for the Propagation of the Gospel in Foreign Parts, the Moravian Mission, the Mission of Bremen, the Methodists, and Roman Catholic missionaries all established themselves on the Gold Coast (Ghana) between 1820 and 1881, opening elementary schools for boys and girls, a seminary, and eventually a secondary school (in 1909). In Nigeria, Protestant missions were opened at Badagry, Abeokuta, Lagos, and Bonny from 1860 to 1899, and the Roman Catholic missions entered afterward and opened the first catechism, primary, secondary, and normal schools. In Uganda and Kenya the Church Missionary Society, the Universities Mission to Central Africa, the White Fathers, and the London Missionary Society opened the first mission schools between 1840 and 1900.

The first official lay schools came later and for a long time constituted a weak minority. In 1899 in Nigeria, for instance, only 33 of the 8,154 primary schools were government-run and only nine of the 136 secondary schools and 13 of the 97 normal schools. Similarly in the Gold Coast in 1914 the government was responsible for only 8 percent of the schools. In Kenya and Uganda, all schools were conducted by missions. Not until 1922 did the British government assume some responsibility for education in Uganda by opening the first government technical school at Makerere (the future Makerere University College). Only in territories seized from the Germans in World War I did the British take over the administration of existing government schools. Generally the British preferred to leave education to missions, which were given variable financial aid, usually from local and inadequate sources.

British efforts at reform in the 1920s and '30s

Following the publication of critical reports in 1922 and 1925, when there was growing uneasiness among the Africans, the missions, the governors, and the administrators, the necessity of a precise policy on education was imposed on the British authorities. In 1925 an Advisory Committee on Education in the Colonies, created in 1924 and presided over by William Ormsby-Gore, published an important report. The ideas, principles, and methods formulated in this document covered the matters involved in defining a policy; namely, the encouragement and control of private educational institutions, the cooperation by the governmental authorities with these institutions, and the adaptation of education to the traditions of the African peoples. Special importance was attached to religious and moral instruction, to the organization and status of education services, to subsidies to private schools, to instruction in the African languages, to the training of native teachers, to the inspection of schools and the upgrading of teachers, to professional training and technique, and to the education of young girls and women. The structure of an educational system, at the most advanced stage, was to consist of an elementary education (generally six years), diversified middle and secondary education (four to six years), technical and professional schools, specialized schools of higher education, and adult education.

In practice, subsequent British policy in black Africa was far from the recommendations of the Ormsby-Gore committee. The subsidies to mission schools were subject to regulations that varied from one colony to another and paid insufficient attention to the character of the education. The development of instruction, especially secondary, was generally curbed, and various local associations and numerous organizations therefore arose to promote the expansion of education. The colonial governments exerted real effort only on behalf of schools that trained subaltern cadres for administration and commerce (mostly schools for the children of chiefs and prominent persons and the colleges at Makerere and Achimota). Government-sponsored secondary education began only after 1930 in the Gold Coast, only in a conditional manner in 1933 at Makerere College in East Africa, and only after 1935 in Nigeria. In Uganda no complete secondary school existed until 1945.

The Advisory Committee reports published in 1935 and 1944 raised the same questions and the same fundamental themes, indicating that the government still was playing an insufficient role in education. Development was primarily a result of the efforts of missions, of various private local or foreign institutions, and of local indigenous authorities. After World War II the different sectors of education were developed with the growing participation of Africans, who were gaining more autonomy. Secondary education expanded. Institutions of higher learning were improved and increased in number: university colleges were established at Accra and Ibadan in 1948, at Makerere in 1949, and at Khartoum in 1951; a College of Technology (later, University of Science and Technology) was founded in Kumasi in 1951; and the Royal Technical College of East Africa (later, University College) was founded in Nairobi in 1954. Beginning in 1950, development plans for the various colonies—Ghana (the Gold Coast), Nigeria, Sierra Leone, Kenya, Uganda, and Tanganyika—contributed to educational progress.

Upon achieving home rule and then independence, the new African states born of the old British colonies were inheritors of an educational system that, though better than that of the other African states, was still a cause for concern. In most states (Ghana, Kenya, and Malawi being the only exceptions) less than 40 percent of the population had a primary education. Secondary education was even less widespread, Ghana being the only country in which it exceeded 10 percent. Higher education existed in urban centres but still in an embryonic state. Other serious obstacles to the ultimate development of education for all the people included the diversity of organizations and institutions responsible for education, the necessity for students to pay fees, and the complexity of the legislation in force.

Educational improvements after independence

Every one of the various countries set out to improve education. They offered subsidies to private schools, extended supervision over them, and regulated their tuition. They increased the number of primary and secondary schools offering free or partly free instruction and created numerous institutions of higher learning, such as the universities of Cape Coast in Ghana, of Lagos, of Ifẹ̀, and of Ahmadu Bello in Nigeria, and the universities of Dar es Salaam in Tanzania, Nairobi in Kenya, and Makerere in Uganda. The educational systems inherited from colonial rule were racially integrated and subjected to "Africanization." The rate of educational growth is not spectacular, however. Moreover, the place made for African languages in primary education seems everywhere to have been eclipsed by English, the official language—in spite of the widespread use of African languages in the mass media.

Education in French colonies and former colonies. As elsewhere in Africa, mission schools were the first to be established in French colonies. Although public or official schools appeared in Senegal between 1847 and 1895, the first such schools in Upper Senegal, Niger, Guinea, the Ivory Coast, and Dahomey were begun only from 1896 on.

Only after 1900, with the organization of the federated colonies of French West Africa and French Equatorial Africa, was there a French colonial policy on education. By decree in 1903, education in French West Africa was organized into a system of primary schools, upper primary schools, professional schools, and a normal school. Two further reorganizations followed decrees in 1912 and 1918, and important schools were established—the St. Louis Normal School in 1907 (transferred to Gorée in 1913), the School for Student Marine Mechanics of Dakar in 1912, and the School of Medicine of Dakar in 1916. The educational organization that remained in force in French West Africa from 1924 until 1947 included a system consisting of primary instruction for six years (regional urban schools), of intermediate–higher primary education given in upper primary schools and in professional schools (generally one for each colony), and at the top the federal schools—two normal schools, a school of medicine and pharmacy, a veterinary school, a school for marine mechanics, and a technical school. The two schools for secondary education, both in Senegal (the Faidherbe State Secondary School of St. Louis and Van Vollenhoven State Secondary School, at Dakar), were reserved for Europeans and those rare Africans having French status.

Total enrollment in French West African schools rose from 15,500 in 1914 to 94,400 in 1945. The number of students in the higher primary schools grew in the same period only from 400 to 800 or 900. (The area's total population in 1945 was almost 16 million.)

French educational policy in Africa

Educational policy was stated frankly in the official statements of governors general:

> Above all else, education proposes to expand the influence of the French language, in order to establish the [French] nationality or culture in Africa (*Bulletin de l'Enseignement en AOF*, No. 45, 1921); Colonial duty and political necessity impose a double task on our education work: on the one hand it is a matter of training an indigenous staff destined to become our assistants throughout the domains, and to assure the ascension of a carefully chosen elite, and on the other hand it is a matter of educating the masses, to bring them nearer to us and to change their way of life. (From *Bulletin de l'Enseignment en AOF*, No. 74, 1931.)

After World War II, all inhabitants of the newly established "French Union" became citizens in common who

were represented in the French Parliament. This political policy carried over into education, which became even more assimilationist: the old higher primary schools, for instance, became classical and modern secondary schools on the French model. An Investment Fund for Economic and Social Development provided financial and developmental aid to education—to the extent that primary enrollments rose to 156,000 in 1950 and to 356,800 in 1957, and higher primary enrollments rose to 5,800 in 1950 and to 14,100 in 1957. Technical and professional education also expanded, from 2,200 students in 1951 to 6,900 in 1957. Scholarships, awarded by the central government, the colonies, and local groups, enabled an increasing number of African youths to pursue higher education in France. In Senegal in 1950 the first French West African university was established, the Institute for Higher Studies, later called the University of Dakar, followed by those of Abidjan and Brazzaville.

In 1957 and 1958, when the colonies achieved autonomy and then a kind of commonwealth status within the new French Community established by the Gaullist constitution, education began a more intensive development, at least quantitatively. More primary and secondary schools were opened; teacher training was accentuated; and more scholarship students went to France. Within three years, after the French African countries had achieved full independence, this upgrading of education accelerated. Curricular reforms, however, have been slow. Although such countries as Guinea, Mali, and the Congo (Brazzaville) introduced such reforms as the Africanization of history and geography, generally the traditional French system persists, and courses are taught in French. The so-called ruralization of primary education—that is, the spread of education out beyond the towns—proceeds under the aegis of the governments and French educational officials.

The rise in the number of primary students was spectacular at first: between 1955 and 1965, for instance, the percentage of primary-age children enrolled in school increased in Guinea from 5 to 31, in Senegal from 14 to 40, in Niger from 2 to 12, and in Chad from 5 to 30. Such progress, however, depended on recourse to unqualified teaching personnel. Since then, some countries have successfully continued programs of rapid educational expansion (the percentage of primary-age children enrolled in school rose to 28 in Niger and to 55 in Senegal in 1985). Progress has been slower in other countries (in 1984 the percentage had risen only to 38 in Chad), and in some areas enrollment has even declined (the percentage dropped to 30 in Guinea in 1985). Also, in the former French areas, the number of students attaining a higher education has remained among the lowest in Africa.

Education in Belgian colonies and former colonies. As elsewhere and perhaps more than elsewhere, the Catholic and Protestant missions played the prime role in the development of education in the Belgian Congo (today Zaire) and in Ruanda-Urundi (the present states of Rwanda and Burundi). In the period before 1908, when the Belgian king Leopold II treated the Congo as virtually his private preserve, the missions had assumed an unofficial responsibility for education. After 1908, when the Belgian parliament assumed control of the Congo, the Roman Catholic mission schools were given a privileged official status, with government subsidies, while the Protestant schools, though financially unassisted, were also officially authorized to operate. Throughout the colonial period the overwhelming majority of schools were missionary, and until 1948 the systems were limited to two-year primary schools, three-year middle schools, and a sprinkling of technical schools for training indigenous cadres. In 1948 the Belgian government issued a new plan entitled "Organization of Free Subsidized Instruction for the Indigenous with the Assistance of Christian Missionary Societies," which promised more diversification in primary education (both vocational and secondary-preparatory) and, more radically, recommended the establishment of secondary schools that would prepare the Congolese for higher education.

Missionary schools in Belgian areas

In 1962 the government of the newly independent Congo proceeded with a reform of the old educational system: the first primary degree was standardized and its length extended from two or four years to six years; and pupils were to take a common primary course prior to their orientation toward general, normal, or technical secondary education or toward professional education. Numerous specialized schools of higher education were created: the National Institute of Public Works and Construction, the School of Civil Engineers, the National Institute of Mines, the Higher Pedagogical Institute, the Higher Institute of Architecture, and the National School of Administration. To the already existing Lovanium University of Kinshasa (founded in 1954) and the Official University of the Congo (founded in 1955 in Lubumbashi) there was added the Free University of the Congo (founded in 1963 in Kinshasa). Although less pronounced, an analogous evolution characterized Burundi and Rwanda.

By 1970 the Congo was among the African countries enjoying the best school and university infrastructure. Then, from 1974 to 1977, the Congo, renamed Zaire in 1971, went through a period of intense nationalism. Zairians with Christian names were ordered to change them to African names. Foreign-owned businesses were sold to Zairian citizens. All schools were nationalized, and mission schools were made state schools. A program of accelerated teacher training was instituted. The old universities were combined with the National University of Zaire. The economic chaos that resulted from these moves caused the government to quickly rescind its plans, however. Businesses were returned to foreign owners in 1977. The church schools were reinstated in 1976, and the universities once again separated in 1981. The result of this upheaval was disastrous for the educational system.

Problems and tasks of education in contemporary Africa. The independent African states face numerous problems in implementing an educational policy that will encourage economic and social development. Pedagogical problems and economic and political problems all intermix. The difficulties confronting most governments, however, are basically political.

African educational crisis

Numerical increases in school enrollments, though occasionally spectacular, fail to correspond to the legitimate aspirations of the people or even, more modestly, to the initial objectives fixed by the governments themselves. The Conference of Nairobi in July 1968 viewed as rather alarming the lack of progress in education and literacy in the context of growing populations. Increasing emphasis has been placed on improving and expanding vocational-technical, adult, and nonformal programs of education. There has also been concern about the financial difficulties of the different states, the unsuitability of current educational systems to local needs, the waste and duplications in primary and secondary education, and the insufficient liaison between educational policy makers and the planners of economic and social development. In short, an educational crisis has developed and ripened in black Africa.

(A.M./Da.G.S.)

The Middle East. Modern education was introduced into the Middle East in the early 19th century through several channels. Rulers in both Egypt and the Ottoman Empire (1300–1922) established new military and civilian schools to teach people the skills required to build modern states. In Iran, too, rulers opened new schools, though on a much smaller scale. Many missionary and foreign schools were also established, especially in the Levant. These modern institutions affected only a small percentage of the people, however; the mass continued to receive a traditional education in the Islāmic schools.

Colonialism and its consequences. Following World War I and the destruction of the Ottoman Empire, new states emerged, which—with the exception of Turkey and Iran—fell under French or British control. Although the new nations inherited educational institutions of various size, each needed to build a new educational system, either from scratch or by expanding a small existing system. Each country sought to use education to provide the skilled manpower required for national development and to socialize its diverse population into feeling loyal to the new state. Educational expansion was pursued everywhere, but the particular pattern of change was profoundly affected by the nature of the political regime, particularly by colonial

status. In Lebanon, Syria, Tunisia, Morocco, and Algeria educational policy reflected French interests. In Egypt, Jordan, Palestine, and Iraq, British policy prevailed. Both colonial powers shared similar goals: to preserve the status quo, train a limited number of mid-level bureaucrats, limit the growth of nationalism, and, especially in the case of France, impose its culture and language. Accordingly, they limited educational expansion, particularly at the higher levels, even though the demand continued to grow.

Aims of colonial educational policy

Private, foreign, and missionary schools were favoured everywhere as alternatives, for the upper classes, to the inadequate public schools. The public systems were centrally administered. Their curricula were usually copied from the British or the French and thus were of limited relevance to local needs; the numbers and quality of teachers were seldom adequate; and dropout rates were high. Few modern schools were to be found in the Arabian Peninsula. Only in Lebanon and in the Jewish community in Palestine (which developed its own educational system) were significant numbers of students enrolled in modern schools. Elsewhere, only a small percentage of the populace (including a few women) received a modern education.

Upon achieving independence, the Middle Eastern nations nationalized the private schools, which were regarded as promoting alien religions and cultures, and greatly expanded educational opportunities, especially at the upper levels. Egypt, for example, in 1925 nationalized a small, poor private institution (founded in Cairo in 1908) and made it into a national university and subsequently opened state universities in Alexandria (1942) and 'Ain Shams (1950). The newly independent countries also sought to equalize educational opportunities. Iraq provided free tuition and scholarships to lower-class students. Syria, in 1946, made primary education free and compulsory. Jordan enacted a series of laws calling for free and compulsory education and placed strict controls on foreign schools, especially the missionary ones.

Despite their importance, these reforms did not transform education. The schools in Egypt, Iraq, Syria, and Jordan, for example, continued to be characterized by rigidity, formalism, high dropout rates, and limited relevance to national needs. Moreover, rapid population increases often offset the educational gains, especially in Egypt. Egypt also could not overcome the existing fragmentation of its educational system. Its modern system was divided into schools for the masses and schools that provided access to the higher levels for the elite. Both types coexisted uneasily with the traditional Islāmic schools, which ran the gamut from rudimentary primary schools to the venerable al-Azhar University.

Educational reforms under Atatürk

Countries with strong nationalist leaders were more successful in modernizing education. Mustafa Kemal Atatürk of Turkey, who was determined to create a modern state, initiated a dramatic program of social and cultural change in which education played an important role. He closed the religious schools, promoted coeducation, prepared new curricula, emphasized vocational and technical education, launched a compulsory adult education project, established the innovative Village Institutes program to train rural teachers, and, in 1933, reorganized Istanbul University into a modern institution staffed mainly by refugees from Nazi Germany. Later, Istanbul Technical University also reorganized and Ankara University was established.

Reza Shah Pahlavi followed similar policies in Iran, albeit to a lesser degree, for he was a reformer rather than a modernizer and ruled a country that had been largely isolated from modern influences. He integrated and centralized the educational system, expanded the schools, especially the higher levels, founded the University of Tehrān (1934), sent students abroad for training, moved against the Islāmic schools, promoted the education of women, and inaugurated an adult education program. Nevertheless, the Iranian educational system remained small and elitist.

After World War II new leaders came to power, including Gamal Abdel Nasser in Egypt in 1952, Habib Bourguiba in Tunisia when it became independent in 1956, and the revolutionary government that deposed the monarchy in Iraq in 1958. They began to make major administrative and social reforms and adopted educational policies similar to those of Atatürk. Bourguiba's reform plans called for universal primary education, an emphasis upon vocational training, expansion of the higher levels, incorporation of the Qur'ānic schools into the modern system, and the promotion of women's education.

Problems of Arabicization in French areas

Tunisia, like the other French possessions in North Africa, had to face yet another educational challenge—nationalizing a system that was designed to socialize students into French culture. Arabicization, the substitution of Arabic for French as the language of instruction and of texts and syllabi representing Arab concerns for ones developed to meet French needs, presented many difficulties. Most teachers were qualified to teach only in French, and appropriate texts were not available. When Algeria and Morocco gained independence from France they adopted similar policies and encountered the same problems, which have been expensive and difficult to overcome. By the 1980s the Arabicization process remained incomplete; in all three countries, some instruction was still being given in French.

Egypt's President Nasser also sought to transform society and culture. He integrated and unified the Egyptian educational system by bringing the religious schools under secular control and by transforming al-Azhar University, long a centre of Islāmic learning, into a modern institution. The old elementary system, which provided access to further education only for urban students, was abolished, and major curricular and other reforms were implemented. All public education was made free, and strong efforts were made to universalize primary education, to upgrade technical and vocational education, and to improve the quality of education generally.

These important reforms did not always produce the anticipated results. Nasser failed to devise a coherent educational strategy that paid adequate attention to the systemic implications and the fit between educational expansion and developments in other sectors. Tunisia, too, despite large investments, was unable to coordinate educational expansion with the needs of the economy.

The contemporary scene. Every modern Middle Eastern state is striving to create an educational system that promotes economic growth and provides equal educational opportunities. In addition, the Arab states wish to promote cultural unity. In 1957 Egypt, Syria, and Jordan replaced the educational structures that had been established by the colonial powers with a common one consisting of six years of primary school, three years of middle school, and three years of secondary school. Most of the Arab states have since followed this pattern, although the length of the school year varies from country to country. The Arab systems also differ in the emphasis they place on certain subjects, especially religious instruction and Arabic, which occupy an especially prominent place in Saudi Arabian schools.

Turkey, Iran, and all the Arab states except Lebanon have another feature in common. Education to the secondary level in these countries is planned and administered by a central ministry. These ministries are generally characterized by administrative weaknesses that severely handicap the provision of education. University education may also be the responsibility of the ministry or, as in Turkey, Iraq, and Egypt, may be supervised by a separate body.

Educational planners have usually attained, or surpassed, their quantitative targets for academic schooling. School enrollments and literacy rates have risen substantially throughout the Middle East. These gains, however, have been at least partly offset by rapidly growing populations. In Egypt the absolute number of illiterates has increased, and Turkey's goal of universalizing primary education was not achieved until the 1980s. Some governments, notably those of Iraq, Algeria, Kuwait, Egypt, Iran, and Turkey, have initiated adult literacy programs, with varying degrees of success.

Inequalities in education

Planners have been less successful in achieving their other goals. Despite great efforts, primary and secondary education everywhere retain certain traditional features. Inequalities remain in such areas as rural and urban access to education and women's education. Although female school enrollment ratios have risen throughout the Mid-

dle East, they remain considerably lower than male ratios in every country except Lebanon and Israel, which have achieved almost universal elementary literacy. Moreover, at the higher levels of education, the percentage of women students becomes progressively lower. Many countries, especially Egypt and Tunisia, have made strenuous efforts to overcome the economic and cultural factors that limit women's education, but their experience demonstrates how difficult it is to do so.

Qualitative goals have also been difficult to achieve. Financial, human, and physical resources have not been able to keep pace with growing enrollments. As a result, the quality of primary and secondary education has suffered. Split shifts, crowded classrooms, serious shortages of qualified teachers, and inadequate textbooks and curricula are common problems. The strict examination system used by most countries to determine which students may advance to the next level of education also hurts educational quality. Most experts agree that the examination system does not provide a valid or reliable indication of student ability. Furthermore, they feel that it reinforces traditional tendencies toward memorization and a rigid classroom culture.

Various innovations have been introduced in an attempt to remedy these shortcomings. One of the most important is the nine-year basic education program, which seeks to provide all children between the ages of six and 15 years with an integrated study program that is practical, does not involve examinations, and prepares students to function in a changing environment. It has been widely implemented in Egypt and has been introduced in Tunisia and Syria as well.

University education in the Middle East

The Middle Eastern nations have also been confronted by serious problems at the university level. Because a degree was widely regarded as a passport to elite status, the demand for higher education grew dramatically. Every government sought to limit the flood of entrants through examinations but managed only to slow, not stop, the growth, which far outpaced projections and resources. To help accommodate the surplus, Egypt and Turkey established programs of "external students" and open universities, which allow students to take courses at home at their convenience through correspondence, radio and television broadcasts, recordings, and other techniques. Every year more and more students of lower-class backgrounds receive a university education, but the entrance examinations tend to limit admissions to the most desirable faculties (medicine and engineering) to students of elite backgrounds. The rising number of graduates with unneeded skills has in turn aggravated problems caused by lack of coordination between education and employment needs. Except in the Gulf states, which have manpower shortages, governments face the difficult task of absorbing poorly prepared graduates into the work force while they try to find qualified managers, technicians and skilled workers.

The development of higher education has been adversely affected by political considerations. Most Middle Eastern countries have never accepted the principle of academic freedom. Turkey, Lebanon, and Israel are prominent exceptions, but even Lebanese and Turkish universities have been subject to political control. The Lebanese civil war has created a difficult environment for education at all levels. In Turkey the universities became so politicized in the 1970s that ideology influenced many aspects of university life. After the military coup of 1980, the government proceeded to limit university autonomy and to eliminate political activism. Iran and most Arab countries have always been ruled by more or less authoritarian regimes that regard universities as potential sources of opposition. The governments in these countries try to use schools and colleges to disseminate the prevailing ideology. Hence scholars often emigrate and those who remain at home are compelled to teach and research in ways that will not create difficulties.

Vocational education

Efforts to increase vocational and technical training have not been very successful because of the continuing appeal of white-collar careers. In Egypt the government's determined attempt to channel students into technical and vocational schools yielded mixed results. Enrollments did increase, but the quality and relevance of such education was questioned as authorities considered the costs involved in purchasing necessary expensive equipment and in training and retaining qualified teachers, whose skills enable them to obtain more remunerative positions in industry. The same difficulties prevail in the other Middle Eastern countries, though Turkey is now initiating, with World Bank support, a promising experiment involving the establishment of a network of modern vocational schools.

Technical training in the agricultural sector is also deficient. There is a shortage of qualified extension agents and other specialists everywhere. Moreover, the bias toward academics means that rural education tends to be neglected, even though the need for agricultural modernization in national development requires that peasants acquire a wide range of skills. Israel is one country where rural education receives the attention that it deserves.

The Islāmic revival. The rapid expansion of modern education and knowledge has produced results that have not been welcomed everywhere. Islām remains, in all societies, a powerful force, one nurtured by traditional factors and by religious education, which continues to be widely offered in one form or another. Believing that traditional Islāmic values have been eroded by Western knowledge based on erroneous assumptions, numerous Islāmic scholars have called for the creation and diffusion of knowledge within an Islāmic framework. The Iranian revolution and the rise of Islāmic fundamentalist movements demonstrate the power of this appeal. Religiously based polities like Saudi Arabia and Iran emphasize Islāmic teachings and values in all schools and colleges, but many other states are providing more religious education than previously.

Migration and the brain drain. Educational systems have also been affected by the widespread international migration of professionals and skilled workers that characterizes the Middle East. Formerly, the West siphoned off a significant percentage of the skilled manpower from Lebanon, Syria, Turkey, Egypt, and Jordan. Now, large numbers of educated persons have migrated from Turkey, Lebanon, Syria, and especially Egypt and Jordan to the oil-rich states, especially to Bahrain, Kuwait, Libya, Saudi Arabia, Qatar, Oman, Algeria, and the United Arab Emirates, all of which face severe manpower shortages. This flow aggravates shortages of skilled workers in many of the exporting countries, especially Jordan, Syria, and Lebanon.

Effects of migration on education and employment

These migration patterns are influenced by and influence educational developments in several ways. They are the result of systems that do not meet a country's labour requirements. The outflows further reduce existing standards, because migrants include the most qualified teachers, especially those with vocational and technical skills. Moreover, the attraction of working abroad is so strong that many persons choose schools and subjects in order to enhance their potential for migration, regardless of the domestic demand. Thus, domestic educational systems have become geared to meet the needs of other societies while domestic employment needs are neglected.

Despite the many problems, it should be emphasized that all the Middle Eastern states have built modern educational systems in the face of considerable difficulties. The importance of education is acknowledged everywhere, and every state is striving to make education more relevant to personal and societal needs, to achieve greater equity, to lower the high wastage rates, and to improve quality.

(J.S.Sz.)

Latin America. The term Latin America is a facile concept hiding complex cultural diversity. This abstraction covers a conglomerate of areas, distinguished by differences not only in the Indian and Negro population base but also in the superimposed nonindigenous patterns—Spanish, Portuguese, French, Dutch, and Anglo-Saxon. In this brief survey, generalizations will be limited to the major Spanish and Portuguese patterns, which account for 95.7 percent of the population and 98.3 percent of the area.

The heritage of independence. At the beginning of the 19th century, the Spanish colonies enjoyed a prosperity that led to optimism, thoughts of independence, and republican rule. In the prolonged struggle for independence,

they were all but ruined, and the change from absolute monarchy to popular democracy was far from easy. The revolutionaries tried to follow the U.S. model, but novel institutions clashed with those of the past; governmental practice did not follow political theory; and the legal equality of the citizens hardly corresponded to economic and educational realities.

Roman Catholicism and education in Latin America

The new governments all considered education essential to the development of good citizens and to the process of modernization. Accordingly, they tried to expand schools and literacy, but they faced two obstacles. Their first was a disagreement over what should form the content of education. Since the time of the Enlightenment, political tyranny and the Roman Catholic church had been blamed for backwardness. Thus, once independence had been achieved, the liberals tried to get rid of the church's privileges and to secularize education. The conservatives, however, wanted to follow traditional educational patterns and considered Catholicism a part of the national character. After decades of confrontation the liberals in many countries managed to make education both secular in character and a state monopoly. In other countries, such as Colombia, by way of a concordat with the Holy See, religious education became the official one.

The second obstacle to educational expansion was a financial one. The new governments lacked the means with which to establish new schools. Thus, they began to import the Lancaster method of "mutual" instruction (so named from its developer, the English educator Joseph Lancaster), which in monitorial fashion employed brighter or more proficient children to teach other children under the direction of an adult master or teacher. Its obvious advantage was that it could accomplish an expansion of education rather quickly and cheaply. Beginning in 1818, it was introduced in Argentina and then in Chile, Colombia, Peru, Mexico, and Brazil. Until well into the second half of the 19th century, it was to be the most widely used system.

Almost all the heroes of independence tried to establish schools and other educational institutions. José de San Martín founded the National Library and the Normal Lancasteriana, a teacher-training school, in Lima; Simón Bolívar established elementary schools in convents and monasteries and founded the Ginecco (1825), known afterward as the Normal Lancasterian School for Women. Bernardino Rivadavia, the first president of Argentina, also stimulated educational development, including the establishment of the University of Buenos Aires. In mid-century, Benito Juárez in Mexico also championed education as the only bulwark against chaos and tyranny.

By the 1870s the liberals had won the day almost everywhere throughout Latin America. Education was declared to be compulsory and free, the lack of teachers and teacher colleges notwithstanding. A program to remedy this situation was launched. Chile paid for the educator Domingo Faustino Sarmiento's travels to the United States and Europe and enabled him to found, on his return in 1842, the Normal School for Teachers. This was the first non-Lancasterian teachers' college and was to be followed in 1850 by the Central Normal School in Lima and in 1853 by the Normal School for Women in Santiago. Countries with more acute educational problems, such as Ecuador, simply imported the Brothers and Sisters of the Sacred Heart and put them in charge of organizing their educational system. During the 1870s and '80s, foreign teachers began to be imported and students were sent abroad. Sarmiento had already called in North American teachers to open his normal schools in the 1860s, and Chile invited Germans for its Pedagogical Institute (1889). Germans and Swiss came to Mexico and Colombia; a number of distinguished Mexican educators were trained by Germans in the Model School in Orizaba. With the foreign professors came new pedagogical ideas—especially those of Friedrich Froebel and Johann Friedrich Herbart—and also new ideologies, foremost among them positivism, which flourished in Argentina, Brazil, Chile, and Mexico.

Administration. With independence, the task of overseeing public instruction fell to the state and local authorities. Fiscal poverty and a lack of trained personnel soon proved them unequal to the task. Furthermore, since most existing schools were confessional and private, the need for intervention by the central authorities to enforce unity became obvious. In 1827 the Venezuelan government established a Subdirectory of Public Instruction, which in 1838 became a directory. Mexico established a General Directory of Primary Instruction in 1833. Soon, some countries decided to assume responsibility for centralization through a ministry for public instruction—Chile and Peru in 1837, Guatemala in 1876, Venezuela in 1881, and Brazil in 1891. Other governments abstained from accepting total responsibility. In Mexico, no ministry was created until 1905 and then only with jurisdiction over the Federal District and territories; even that became a victim of the revolution of 1910. In 1922 a Mexican ministry was reestablished, now in charge of the whole republic and taking up the functions that the states could not fulfill. In Argentina the Lainez Law, decreed in 1905, authorized the National Council of Education to maintain, if need be, schools in the provinces.

Efforts to centralize educational administration

Today, in all countries the control over education is in the hands of a ministry of public education or a similar government unit. Its functions include planning, building, and administering schools, authorizing curricula and textbooks for public elementary and secondary schools, and supervising private ones. In some countries, the states sustain their own educational systems, which the federal government then supplements, but, because of the disparity between city and countryside, these federal governments often have had to shoulder almost the total burden of rural elementary education.

Primary education and literacy. At the time of independence, elementary education consisted of teaching reading and writing, the religious and civil catechisms, and rudiments of arithmetic and geometry. By the second half of the century, it became differentiated between "elementary primary" and "superior primary" education, and the curriculum was enlarged to include the teaching of national language, history, geography, rudimentary natural sciences, hygiene, civics, drawing, physical education, and crafts for boys and needlework for girls. The elementary primary school was increased to five or six years, and the superior primary was to become the secondary school of the 20th century. These educational levels absorbed the greatest part of the governmental efforts and became a means to do away with illiteracy and also to create a concept of citizenship.

Primary instruction was improved by special programs and teacher training, and both benefited from educational influences coming from abroad but also from improvements resulting from the study of national problems. Today, primary-school teachers are trained in teachers' colleges having the status of secondary schools.

Problems of rural and native education

Thanks to solid foundations laid during the 19th century, public education in Argentina and Chile reached a high level of competence. In other countries, because of such factors as a more heterogeneous population, a higher level of demographic growth, and greater geographical barriers, the results of great efforts have been less than impressive. Although all countries have declared primary instruction to be free and compulsory, the situation in reality is rather complex. Whereas, in towns, many children have gone from kindergarten to secondary schools since the beginning of the century, in the rural areas, even today, many schools have only one teacher to handle students of all levels. Furthermore, because many Indian citizens do not understand Spanish, special instruction is required. In the 20th century, governments have established special institutions for Indians. The first such cultural mission was created by the Mexican secretary of education, José Vasconcelos, in 1923. The idea was to send an elementary-school teacher, an expert in trades and crafts, a nurse, and a physical-education teacher to underdeveloped communities, in which, during a limited period, the population would be provided with some general education. The United Nations Educational, Scientific and Cultural Organization (UNESCO) has helped in the training of teachers for these special areas through two regional centres of fundamental education for Latin America (CREFAL), one in

Mexico and the other in Venezuela. Many countries have tried to master the dropout problem by offering at least one free meal a day to those who continue their schooling.

Uruguay, Argentina, and Chile have been able to multiply their schools and thus to provide facilities for their entire population of school age. In other countries, the efforts may be gauged by comparing statistics. In Peru, only 29,900 children went to school in 1845; but there were 59,000 in 1890 and 2,054,000 in 1965. In Brazil, there were 115,000 pupils in 1869; 300,000 in 1889; and 9,923,000 in 1965. In Mexico, there were 349,000 in 1874; 800,000 in 1895; and 7,813,000 in 1969. Unfortunately, the high population-growth rate (2.9 percent) makes it difficult to keep up with the ever-increasing needs.

Fight against illiteracy

Illiteracy has been fought by various means in accordance with the political and socioeconomic situation. Until the middle of the 19th century, illiteracy in Latin America was in excess of 90 percent. Of Brazil's population, only 1.5 percent were literate in 1823. Around the beginning of the 20th century, illiteracy had decreased to 39 percent in Argentina (1908), 50.4 percent in Uruguay (1908), and 68.2 percent in Chile (1895); in other countries it fluctuated between 80 and 98 percent. By 1985 illiteracy was down to 6.0 percent in Argentina, 5.7 percent in Uruguay, 10 percent in Chile, 26 percent in Mexico, 28 percent in Peru, 25 percent in Bolivia, and 26 percent in Brazil. Nations with the greatest illiteracy were Guatemala, with 50 percent, and Haiti, with 77 percent.

Secondary education. During the 19th century, many countries established new secondary schools on the basis of colonial institutions. Thus, in 1821 Argentina converted its College of San Carlos into its College of Moral Sciences. Mexico attempted a total reform in 1833 but would not complete it until 1867 with the founding of the National Preparatory School, which involved reforming the whole system on the basis of positivist philosophy. In Brazil the Royal Military Academy was established in 1810 and the Pedro II College in 1830, but secondary instruction did not prosper until the return of the Jesuits in 1845 and was to be supplemented later by *gimnasios*—that is, *Gymnasien* on the German model. Peru and Venezuela established national colleges, and Chile and Argentina created *liceos* (modeled on the French *lycées*) and, later, national colleges. (The term college in all cases here is used in the continental European sense to refer to secondary institutions, not institutions of higher education.)

Secondary emphasis on university preparation

In all countries (except perhaps Chile), secondary instruction has been considered a preparation for the university. All attempts to make it more formative and practical have failed, in spite of the fact that the government has taken charge. The secondary-preparatory course lasts from five to six years, with a degree of bachelor (*bachillerato*) usually given upon its completion. Its teachers come from the humanities departments of the universities and the superior normal schools (which have existed since 1869 in Argentina, since 1889 in Chile, and in the 20th century in the other countries).

Polytechnical education—industrial, commercial, and agricultural—had been a concern of liberal governments since the end of the 19th century but has developed only recently. Traditional prejudices against practical instruction were overcome only after industrialization began. It has been emphasized only in Argentina, Venezuela, Chile, and Mexico.

Higher education. Imbued with a revolutionary spirit in which education was a vital element, Latin Americans founded 10 universities between 1821 and 1833, among them the University of Buenos Aires (1821). Bolívar himself established two in Peru—Trujillo (1824) and Arequipa (1828). With independence, practically all theological faculties had disappeared, and their position of preeminence was taken over by faculties of law.

Four universities were founded in the 1840s, Chile's among them, and 10 more in the second half of the 19th century. In Mexico the new institutions called themselves institutes of arts and sciences, because the University of Mexico (founded in 1551) was associated with colonialism and had become a favourite target of the liberals. The University of Mexico was suppressed in 1865, not to be reopened until 1910, the year of the revolution. Argentine liberals solved their problem by passing the Avellaneda Law (1885), which allowed only national universities, prohibiting private universities (until the reform of 1955).

In Brazil the plans to open a university in 1823 failed. Several professional schools were established, but the first university opened its doors in 1912 in Paraná. In 1920 the Federal University of Rio de Janeiro was founded.

Almost all higher education in Latin America came to be secular and state-operated. The fact that Latin-American governments, themselves unstable, generally took charge of higher education, however, explains in part its uncertain existence.

Emphasis on state universities

Some colonial religious institutions nevertheless survived. During part of the 19th century, for instance, the University of the Republic in Montevideo maintained its ties with the church. In 1855 the University of San Carlos in Guatemala, through a concordat with Rome, reverted to pontifical status. But, with the exception of the Catholic University in Chile (1888), the Pontifical Catholic University of Peru (1917), and the Javeriana University in Colombia (1931), all religious universities are recent creations. Indeed, today the majority of private institutions are religious or confessional, with the significant exception of some recently established technological institutes (in Monterrey, Mex., and in Buenos Aires). The need for technical education was also recognized by the Mexican government when it founded, in 1936, the National Polytechnical Institute as its second national institution of higher learning, with several branches in the country (regional technological institutes) to serve the particular needs of each region.

Until the 20th century, universities were mainly professional schools. Often, they also supervised primary and secondary education (Uruguay, 1833–37; Chile, 1842–47; Mexico, 1917–21). Today, they also conduct research and try to encourage regional developments. Unofficially, they have sometimes played a role in political life. Since the reform movement for student representation at the University of Córdoba in Argentina in 1918, they have become involved in political controversies. The Mexican government tried to extricate the National University from political strife by giving it autonomy in 1929. Student demonstrations by the late 1960s, however, proved this measure to lack effectiveness.

Higher education has proved to be the best means of furthering social mobility. In spite of this, institutions of higher learning in Latin America have suffered from several handicaps. Foremost is the lack of sufficient funds, which usually results in poor research facilities. Second, both students and professors are generally engaged only half-time. This increases the dropout rate and decreases performance. Thus, most highly qualified professionals are trained abroad. At the same time, both the political situation and economic pressures have induced an exodus of the most highly educated Latin Americans to the United States.

In 1985 there were more than 1,500 institutions of higher learning in Latin America. Brazil, Mexico, Argentina, and Colombia had the highest numbers of university students, but, on the basis of the number of students per population, Argentina was first, distantly followed by Uruguay, Cuba, Chile, Colombia, Mexico, and Brazil.

Reform trends. Although most of the Latin-American countries achieved nominal independence in the 19th century, they remained politically, economically, and culturally dependent on U.S. and European powers throughout the first half of the 20th century. By 1960, many viewed this dependency as the reason for Latin America's state of "underdevelopment" and felt that the situation could best be remedied through educational reform. The most general reform movement (*desarrollista*) simply accepted the idea of achieving change through "modernization," in order to make the system more efficient. The Brazilian educationist Paulo Freire, however, advocated mental liberation through self-consciousness, a view that was influential in the 1960s and '70s throughout Latin America. Because political dictatorship prevailed through the 1960s and part of the 1970s in many countries, authoritarian

pedagogy became the practice, especially in Chile. In the 1980s the deep economic crisis in Latin America proved to be the greatest influence on education, obstructing all renovation or modernization of public education.

(J.Z.V.)

Historical background of modern Southeast Asian education

Southeast Asia. Indigenous culture, colonialism, and the post-World War II era of political independence influenced the forms of education in the nations of Southeast Asia—Myanmar (Burma), Kampuchea (Cambodia), Indonesia, Laos, Malaysia, the Philippines, Singapore, Thailand, and Vietnam.

Before AD 1500, education throughout the region consisted chiefly of the transmission of cultural values through family and community living, supplemented by some formal teaching of each locality's dominant religion—animism, Hinduism, Buddhism, Taoism, Confucianism, or Islām. Religious schools typically were attended by boys living in humble quarters at the residence of a pundit who guided their study of the scriptures for an indeterminate period of time.

With the advent of Western colonization after 1500, and particularly from the early 19th to mid-20th century, Western schooling with its dominantly secular curriculum, sequence of grades, examinations, set calendar, and diplomas began to make strong inroads on the region's traditional educational practices. For the indigenous peoples, Western schooling had the appeal of leading to employment in the colonial government and in business and trading firms.

After World War II, as all sectors of Southeast Asia gained political independence, each newly formed nation attempted to achieve planned development—to furnish primary schooling for everyone, extend the amount and quality of postprimary education, and shift the emphasis in secondary and tertiary education from liberal, general studies to scientific and technical education. Although indigenous culture was still learned through family living and traditional religion continued to be important in people's lives, most formal schooling throughout Southeast Asia had become predominantly of a Western, secular variety.

Schooling in all of these countries was organized in three main levels, primary, secondary, and higher. In addition, nursery schools and kindergartens, operated chiefly by private groups, were gradually gaining popularity. The typical length of primary schooling was six years. Secondary education was usually divided into two three-year levels. A wide variety of postsecondary institutions offered academic and vocational specializations. Beginning in the 1950s, nonformal education to extend literacy and vocational skills among the adult population expanded dramatically throughout the region. Most of the nations were committed to compulsory basic education, typically for six years but up to nine years in Vietnam. However, by the close of the 1980s, the inability of governments to furnish enough schools for their growing populations prevented most from fully realizing the goal of universal basic schooling.

In each nation a central ministry of education set schooling structures and curriculum requirements, with some responsibilities for school supervision, curriculum, and finance often delegated to provincial and local educational authorities. Government-sponsored educational research and development bureaus had been established since the 1950s in an effort to make the countries more self-reliant in fashioning education to their needs. Regional cooperation in attacking educational problems was furthered by membership in such alliances as the Southeast Asian Ministers of Education Organization (SEAMEO) and the Association of Southeast Asian Nations (ASEAN).

Problems which most Southeast Asian education systems continued to face were those of reducing school dropout and grade-repeater rates, providing enough school buildings and teachers to serve rapidly expanding numbers of children, furnishing educational opportunities to rural areas, and organizing curricula and the access to education in ways that suited the cultural and geographical conditions of multiethnic populations.

Myanmar (formerly Burma). The indigenous system of education in Myanmar consisted mainly of Buddhist monastic schools of both primary and higher levels. They were based on (1) the moral code of Buddhism, (2) the divine authority of the kings, (3) the institution of *myothugyi* (township headmen), and (4) widespread male literacy. The Western system was established after the British occupation in 1886. The new system recognized women's right to formal education in public schools, and women began to play an increasingly important role as teachers. The Government College at Rangoon and the Judson College established in the 19th century were incorporated as the University of Rangoon under the University Act of 1920.

Buddhist and Socialist influence in Myanmar

Following independence in 1948, the country experienced more than a decade of political instability until a coup d'état in 1962 brought a strongly centralized socialist government to power. Subsequently, marked improvements in education occurred. Science was emphasized along with general academic subjects, civic education, and practical arts. Primary-school attendance for children ages five through nine became free where available. From 1965 to 1985 enrollments increased in primary schools from two to five million, in secondary schools from 503,000 to 1.25 million, and in higher education from 21,000 to 189,000.

Malaysia and Singapore. The Malay States, Singapore, and sectors of North Borneo were British colonies until reorganized as the nation of Malaysia in 1963. Singapore left the coalition in 1965 to become an independent city-nation. As a result, while Malaysia and Singapore share common educational roots, their systems have diverged since 1965.

The language issue in Malaysia and Singapore

Under British rule, the most significant feature of education on the Malay peninsula was the structuring of primary schools in four language streams—Malay, Chinese, English, and Tamil. Students in the English stream enjoyed favoured access to secondary and higher education as well as to employment in government and commerce. After 1963 Malaysian leaders sought to indigenize and unify their society by adopting the Malay language as the medium of instruction in schools beyond the primary level and by teaching English only as a second language. In contrast, the government of Singapore urged everyone to learn English, plus one other local tongue—Chinese, Malay, or Tamil. Thus, in both nations the learning of languages became a critical issue in people's efforts to gain access to socioeconomic opportunity and in political leaders' attempts to unify their multiethnic populations.

Efforts to popularize schooling in Malaysia and Singapore were notably successful. By the early 1980s, 93 percent of all Malaysian children ages six to 11 attended primary school, with nearly 90 percent of primary-school graduates entering lower-secondary school. By 1968, all primary-age children in Singapore were in school. In both countries, secondary- and higher-education enrollments continued to increase rapidly. Both nations were well supplied with school buildings, textbooks, and trained teachers.

Indonesia. From AD 100 to 1500 the Indonesian aristocracy adopted Hindu and Buddhist teachings, while education for the common people was provided mainly informally, through daily family living. Islām, introduced into the archipelago around 1300, spread rapidly in the form of Qur'ān schools, which have continued through the 20th century, though in diminishing numbers. The first few schools on Western lines were established by Portuguese and Spanish priests in the 16th century. As the Dutch colonialists gained increasing control over the islands, they set up schools patterned after those in Holland, primarily for European and Eurasian pupils. In 1848 the Dutch East Indies government officially committed itself to providing education for the native population. However, even though the amount of education for indigenous islanders increased over the following century, Western schooling under the Dutch never reached the majority of the population.

After Indonesians gained independence from the Dutch in 1949, they sought to provide universal elementary schooling and a large measure of secondary and higher education. Progress toward this goal after 1950 was rapid, despite the challenge of an annual population growth rate of around 2.3 percent. Enrollments over the 1950–1985 period increased from five million to 30 million at the

elementary level, from 230,000 to 7.5 million at the secondary level, and from 6,000 to one million at the tertiary level. Although the Indonesian population was 90 percent Muslim, three-fourths of the nation's schools were of a Western secular variety. The remaining one-fourth were Islāmic schools required to offer at least 70 percent secular studies and no more than 30 percent religious subjects. This ratio reflected the government's efforts to use the schools for preparing manpower for socioeconomic modernization, as guided by a sequence of five-year national development plans.

Philippines. The pre-Spanish Philippines possessed a system of writing similar to Arabic, and it was not uncommon for adults to know how to read and write. Inculcation of reverence for the god Bathala, obedience to authority, loyalty to the family or clan, and respect for truth and righteousness were the chief aims of education. After the Spanish conquest, apart from parochial schools run by missionaries, the first educational institutions to be established on Western lines were in higher education. The Santo Tomás College, established in 1611 and raised to the status of a university in 1644–45, served for centuries as a centre of intellectual strength to the Filipino people. Educational growth, however, was slow, mainly because of lack of government support.

Spanish and American influences in the Philippines

With the advent of American rule, the stress laid on universal primary education in the policy announced by U.S. President William McKinley on April 7, 1900, led to a rapid growth in primary education. A number of institutions of higher education were also established between 1907 and 1941, including the University of the Philippines (1908). Private institutions of higher education, however, far outnumbered the state institutions, thus indicating a trend that remains a characteristic feature of the system of higher education in the Philippines.

The new Republic of the Philippines emerging after World War II launched a series of national development plans that included components aimed at the renovation and expansion of education to promote socioeconomic modernization. Over the period 1948 to 1986, enrollments rose in primary schools from four million to nine million and in secondary schools from 424,000 to 3.3 million. By the late 1980s, 1.5 million students were in the nation's more than 1,000 higher-education institutions. More than 95 percent of primary pupils and 41 percent of secondary students attended public schools, while the remainder attended private institutions.

Thailand. The traditional system of education in Thailand was inspired by the Thai philosophy of life based on (1) dedication to Theravāda Buddhism, with its emphasis on moral excellence, generosity, and moderation, (2) veneration for the king, and (3) loyalty to the family. The beginning of the present system of education can be traced to 1887, when King Chulalongkorn set up a department of education with foreign advisers, mostly English educationists. Gradually, temple schools were established. The process of westernization of education was strengthened with the establishment of a medical school in 1888, a law school in 1897, and a royal pages' school in 1902 for the general education of "the sons of the nobility." It was converted into the Civil Service College in 1910.

The Thai philosophy

The abolition of the absolute monarchy after the 1932 revolution stimulated the government to increase educational provisions at all levels, particularly for training specialists in higher-learning institutions. Beginning in 1962, the nation's series of five-year development plans assigned educational institutions a crucial role in manpower preparation. The government supervises all educational institutions, public and private. Financing education is primarily a government responsibility, supplemented by the private sector. Thai is the language of instruction at all levels, with English taught as a second language above grade four.

By the mid-1980s there were more than 7.3 million pupils (over 90 percent of the age group) enrolled in the compulsory six-year elementary schools, 2.2 million in the six years of secondary schooling, and 715,000 in the nation's 31 registered universities and colleges.

Kampuchea (formerly Cambodia). For nearly four centuries before the advent of the French in 1863, the educational system in Cambodia grew up around Theravāda Buddhism, which became the established religion toward the end of 1430 under Thai influence. In 1887 Cambodia became a part of the French Indochina Union and did not achieve complete independence until 1954. Pagoda schools, imparting education at the primary level, were remodeled and integrated into the primary school system administered by the Ministry of Education.

Indo-Chinese education under colonialism and war

Civil war throughout the 1970s disrupted education until Vietnamese forces overthrew the Khmer Rouge government in 1979. By the mid-1980s schools had reopened with a total enrollment of nearly two million throughout the four-year primary, three-year junior-secondary, and three-year senior-secondary structure. Secondary schools and the country's few higher-education colleges were in a state of rebuilding. Much of the teacher-training was in the form of short courses, and nonformal adult literacy classes multiplied at a rapid pace.

Laos. The pagoda school was the main unit of the traditional educational system in Laos. Efforts toward modernization came in the wake of the country's becoming a French protectorate in 1893 and finally after its inclusion in 1904 within the French Indochina Union. The medium of education was changed to French when the French Education Service was created.

In 1975, after 30 years of uninterrupted revolution, a socialist government was established and schooling was accorded high priority. By the mid-1980s 79 percent of all children seven to 11 years old were in the five-year primary school, 48 percent of children 12 to 14 years old were in the three-year junior-secondary school, and 23 percent of the 15- to 17-year-olds were in the three-year senior-secondary school.

Vietnam. Long Chinese domination over the emperors of Vietnam resulted in strong Confucian and Taoist influences on the Vietnamese educational system, though it centred on Buddhism. The establishment of French rule, commencing with the occupation of Saigon (now Ho Chi Minh City) in 1859, led to the gradual growth of a pattern of education similar to that of the rest of the former Indochina Union. Vietnamese attempts to develop education were thwarted by the continued fighting from World War II onward and, after the partition of the country in 1954, by fighting between the South and the North. After the war's end in 1975, the Communist government attempted to "reeducate" the conquered South and sought to establish urgently needed technical and vocational education in secondary and higher levels. By the mid-1980s there were eight million pupils in elementary schools, four million in secondary schools, and more than 115,000 in higher-education institutions. (M.S.H./R.M.T.)

BIBLIOGRAPHY

General works: General histories of education are mainly concerned with the educational history of the West. In some works early chapters survey non-Western educational developments in the context of ancient civilizations, and medieval Muslim education is frequently treated because of its impact upon Western education. Given these limitations, among the best general histories are ELLWOOD P. CUBBERLEY, *The History of Education* (1920, reissued 1948); JAMES BOWEN, *A History of Western Education,* 3 vol. (1972–81); WILLIAM BOYD and EDMUND J. KING, *The History of Western Education,* 11th ed. (1975, reprinted 1980); R. FREEMAN BUTTS, *The Education of the West* (1973); ROBERT ULICH, *The Education of Nations,* rev. ed. (1967), and *History of Educational Thought,* rev. ed. (1968); HARRY G. GOOD and JAMES D. TELLER, *A History of Western Education,* 3rd ed. (1969); JAMES MULHERN, *A History of Education: A Social Interpretation,* 2nd ed. (1959); MEHDI NAKOSTEEN, *The History and Philosophy of Education* (1965); and MARGARET SCOTFORD ARCHER, *Social Origins of Educational Systems* (1979).

Despite its age, The five-volume *A Cyclopedia of Education,* ed. by the American educator PAUL MONROE (1911–13, reprinted 1968), remains a comprehensive source of historical information. Its influence was recognized in FOSTER WATSON (ed.), *The Encyclopaedia and Dictionary of Education,* 4 vol. (1921–22), a British work whose foreign contributors included John Dewey and Benedetto Croce. LEE C. DEIGHTON (ed.), *The Encyclopedia of Education,* 10 vol. (1971), also has numerous historical references. There are many national encyclopaedias of historical interest in education.

Among historical surveys of individual countries, the fol-

lowing are useful: W.H.G. ARMYTAGE, *Four Hundred Years of English Education,* 2nd ed. (1970); S.J. CURTIS, *History of Education in Great Britain,* 7th ed. (1967); CHRISTOPHER BROOKE and ROGER HIGHFIELD, *Oxford and Cambridge* (1988); CHARLES FOURRIER, *L'Enseignement français de l'Antiquité à la Révolution* (1964), and *L'Enseignement français de 1789 à 1945* (1965), on France; WILLIAM H.E. JOHNSON, *Russia's Educational Heritage* (1950, reissued 1969); TOKIOMI KAIGO, *Japanese Education: Its Past and Present,* 2nd ed. (1968); PING-WEN KUO, *The Chinese System of Public Education* (1915, reprinted 1972); T.N. SIQUEIRA, *The Education of India: History and Problems,* 4th rev. ed. (1952); AHMAD SHALABY, *History of Muslim Education* (1954, reissued 1979); ALLAN BARCAN, *A History of Australian Education* (1980); ROGER OPENSHAW and DAVID MCKENZIE (eds.), *Reinterpreting the Educational Past: Essays in the History of New Zealand Education* (1987); LAWRENCE A. CREMIN, *American Education, the Colonial Experience, 1607–1783* (1970), *American Education, the National Experience, 1783–1876* (1980), and *American Education, the Metropolitan Experience, 1875–1980* (1988); DAVID B. TYACK, *The One Best System: A History of American Urban Education* (1974); and J. DONALD WILSON, ROBERT M. STAMP, and LOUIS-PHILIPPE AUDET (eds.), *Canadian Education* (1970).

Education in primitive and early civilized cultures: There are few monographs dealing solely with education in primitive civilizations; information is to be found chiefly in works treating larger subjects, such as MARGARET MEAD, *Continuities in Cultural Evolution* (1964); GEORGE DEARBORN SPINDLER (ed.), *Education and Cultural Process: Anthropological Approaches,* 2nd ed. (1987); THOMAS WOODY, *Life and Education in Early Societies* (1949, reprinted 1970); CHRISTOPHER J. LUCAS, *Our Western Educational Heritage* (1971); HENRI MASPERO, *China in Antiquity* (1978; originally published in French, 1927); J. ERIC S. THOMPSON, *The Rise and Fall of Maya Civilization,* 2nd enlarged ed. (1966, reprinted 1977); RUDOLPH VAN ZANTWIJK, *The Aztec Arrangement: The Social History of Pre-Spanish Mexico* (1985; originally published in Dutch, 1977); and GEORGE A. COLLIER, RENATO I. ROSALDO, and JOHN D. WIRTH (eds.), *The Inca and Aztec States, 1400–1800* (1982).

Education in classical cultures: In addition to the treatments offered in the general histories cited above, see HOWARD S. GALT, *A History of Chinese Educational Institutions: To the End of the Five Dynasties, A.D. 960* (1951); FREDERICK A.G. BECK, *Greek Education, 450–350 B.C.* (1964), and *Album of Greek Education: The Greeks at School and at Play* (1975); STANLEY F. BONNER, *Education in Ancient Rome: From the Elder Cato to the Younger Pliny* (1977); M.L. CLARKE, *Higher Education in the Ancient World* (1971); JOHN P. LYNCH, *Aristotle's School: A Study of a Greek Educational Institution* (1972); O.W. REINMUTH, *The Ephebic Inscriptions of the Fourth Century B.C.* (1971); W.H. STAHL, R. JOHNSON, and E.L. BURGE, *Martianus Capella and the Seven Liberal Arts* (1971); RADHAKUMUD MOOKERJI, *Ancient Indian Education: Brahmanical and Buddhist,* 4th ed. (1969); and NATHAN DRAZIN, *History of Jewish Education from 515 B.C.E. to 220 C.E.* (1940, reprinted 1979).

Education in Persian, Byzantine, early Russian, and Islāmic civilizations: Ancient Persian culture and civilization are studied in MANECKJI NUSSERVANJI DHALLA, *Zoroastrian Civilization* (1922, reprinted 1977). For surveys of Byzantine education, see appropriate chapters in STEVEN RUNCIMAN, *Byzantine Civilization* (1933, reissued 1975); and NORMAN H. BAYNES and HENRY ST. L.B. MOSS (eds.), *Byzantium: An Introduction to East Roman Civilization* (1948, reprinted 1969). Special works include PAUL LEMERLE, *Byzantine Humanism, the First Phase: Notes and Remarks on Education and Culture in Byzantium from Its Origins to the 10th Century* (1986; originally published in French, 1971); and N.G. WILSON, *Scholars of Byzantium* (1983). On early Russian education, see NICHOLAS HANS, *The Russian Tradition in Education* (1963, reprinted 1973); WILLIAM K. MEDLIN and CHRISTOS G. PATRINELIS, *Renaissance Influences and Religious Reforms in Russia: Western and Post-Byzantine Impacts on Culture and Education, 16th–17th Centuries* (1971); and HUGH F. GRAHAM, "Did Institutionalized Education Exist in Pre-Petrine Russia?" in DON KARL ROWNEY and G. EDWARD ORCHARD (eds.), *Russian and Slavic History* (1977), pp. 260–273. Medieval Muslim education and its impact upon Western education is studied in GEORGE MAKDISI, *The Rise of the Colleges: Institutions of Learning in Islam and the West* (1981), an authoritative work; and MEHDI NAKOSTEEN, *History of Islamic Origins of Western Education, A.D. 800–1350* (1964).

The European Middle Ages: Some of the best surveys of medieval European education are contained in the general histories of education listed at the beginning of this bibliography. On elementary and grammar schooling of the period, the first major work was A.F. LEACH, *The Schools of Medieval England* (1915, reprinted 1969). Also important are JOAN SIMON, *Education and Society in Tudor England* (1966, reprinted 1979), which also covers the Renaissance and the Reformation; JOHN WILLIAM ADAMSON, *The Illiterate Anglo-Saxon: And Other Essays on Education, Medieval and Modern* (1946, reprinted 1977); and NICHOLAS ORME, *English Schools in the Middle Ages* (1973). For higher learning, see R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954, reprinted 1977); CHARLES HOMER HASKINS, *The Rise of Universities* (1923, reprinted 1976); HASTINGS RASHDALL, *The Universities of Europe in the Middle Ages,* new ed., ed. by F.M. POWICKE and A.B. EMDEN, 3 vol. (1936, reprinted 1987), a standard work; HELENE WIERUSZOWSKI, *The Medieval University: Masters, Students, Learning* (1966); and ALAN B. COBBAN, *The Medieval Universities: Their Development and Organization* (1975). Relevant monographs are WILLIAM J. COURTENAY, *Schools & Scholars in Fourteenth-Century England* (1987); DAVID KNOWLES, *The Evolution of Medieval Thought,* 2nd ed. (1988); and NANCY G. SIRAISI, *Arts and Sciences at Padua: The Studium of Padua Before 1350* (1973).

Education in Asian civilizations, c. 700 to the eve of Western influence: S.M. JAFFAR, *Education in Muslim India* (1936, reprinted 1973), is a vivid documentary account. NARENDRA NATH LAW, *Promotion of Learning in India During Muhammadan Rule, by Muhammadans* (1916, reprinted 1984 with a new introduction), is informative. For China and Japan, see EDWARD A. KRACKE, *Civil Service in Early Sung China, 960–1067* (1953, reprinted 1968); R.P. DORE, *Education in Tokugawa Japan* (1965, reprinted 1984); and RICHARD RUBINGER, *Private Academies of Tokugawa Japan* (1982).

European Renaissance and Reformation: Introductions to Renaissance education include WILLIAM HARRISON WOODWARD, *Studies in Education During the Age of the Renaissance, 1400–1600* (1906, reprinted 1967), *Vittorino da Feltre and other Humanist Educators* (1897, reprinted 1970), and *Desiderius Erasmus Concerning the Aim and Method of Education* (1904, reprinted 1971). See also DAVID CRESSY, *Literacy and the Social Order: Reading and Writing in Tudor and Stuart England* (1980). Important works on the Reformation and Counter-Reformation are JOHN LAWSON, *Mediaeval Education and the Reformation* (1967); FREDERICK EBY, *Early Protestant Educators: The Educational Writings of Martin Luther, John Calvin, and Other Leaders of Protestant Thought* (1931, reprinted 1971); GERALD STRAUSS, *Luther's House of Learning* (1978); and ALLAN P. FARRELL, *The Jesuit Code of Liberal Education* (1938).

European education in the 17th and 18th centuries: The general histories cited at the beginning of this bibliography offer good accounts of educational developments of the 17th and 18th centuries. For the 17th century, a useful work is JOHN WILLIAM ADAMSON, *Pioneers of Modern Education 1600–1700* (1905, reissued 1972). Major theorists are treated in JEAN PIAGET, "Introduction," in JOHN AMOS COMENIUS, *Selections* (1957), published by UNESCO; JOHN W. YOLTON, *John Locke & Education* (1971); MICHAEL MOONEY, *Vico in the Tradition of Rhetoric* (1985); H.C. BARNARD, *The French Tradition in Education: Ramus to Mme. Necker de Saussure* (1922, reprinted 1970); WILLIAM BOYD, *The Educational Theory of Jean Jacques Rousseau* (1911, reissued 1963); ALLAN BLOOM, "Introduction," in his edition of JEAN JACQUES ROUSSEAU, *Emile: Or, On Education* (1979); and J.J. CHAMBLISS, *Educational Theory as Theory of Conduct: From Aristotle to Dewey* (1987). Introductions to the 18th century include NICHOLAS HANS, *New Trends in Education in the Eighteenth Century* (1951, reprinted 1966); F. DE LA FONTAINERIE (ed.), *French Liberalism and Education in the Eighteenth Century: The Writings of La Chalotais, Turgot, Diderot, and Condorcet on National Education* (1932, reprinted 1971); and L.W.B. BROCKLISS, *French Higher Education in the Seventeenth and Eighteenth Centuries* (1987). For European influence on colonial developments, see LUÍS MARTÍN and JO ANN GEURIN PETTUS (eds.), *Scholars and Schools in Colonial Peru* (1973); and JOSEPH MAIER and RICHARD W. WEATHERHEAD, *The Latin American University* (1979).

Western education in the 19th century: This period is treated in the general histories cited above. *The American Journal of Education* (1856–82), ed. by HENRY BARNARD, remains a valuable source for European and U.S. educational developments. For analysis of theories, see KATE SILBER, *Pestalozzi: The Man and His Work,* 3rd ed. (1973); and JOHN ANGUS MACVANNEL, *The Educational Theories of Herbart and Froebel* (1905, reissued 1972). Works on individual countries include FRIEDRICH PAULSEN, *German Education Past and Present* (1908, reprinted 1976; originally published in German, 1906), a classic analysis; JOHN WILLIAM ADAMSON, *English Education, 1789–1902* (1930, reprinted 1964); PATRICK L. ALSTON, *Education and the State in Tsarist Russia* (1969); BEN EKLOF, *Russian Peasant Schools* (1986); BRUCE CURTIS, *Building the Educational State: Canada West, 1836–1871* (1988); A.G. AUSTIN, *Australian Education, 1788–1900,* 2nd ed. (1965); and A.G. BUTCHERS, *Young New Zealand: A History of the Early Contact of the Maori Race with the European, and of the Establishment of a National System of Education for Both Races* (1929). The spread of Western influences to Asia is studied in MAKOTO ASO and IKUO AMANO,

Education and Japan's Modernization (1972, reissued 1983); SYED NURULLAH and J.P. NAIK, *A History of Education in India During the British Period,* 2nd rev. ed. (1951, reissued 1968); S.N. MUKERJI, *History of Education in India: Modern Period,* 6th ed. (1974); and BHAGWAN DAYAL SRIVASTAVA, *The Development of Modern Indian Education,* rev. ed. (1963).

Education in the 20th century: Surveys of 20th-century practices and theories are found in the general histories listed at the beginning of this bibliography. See also ROBIN BARROW and GEOFFREY MILBURN, *A Critical Dictionary of Educational Concepts* (1986); T. NEVILLE POSTLETHWAITE (ed.), *The Encyclopedia of Comparative Education and National Systems of Education* (1988); J. CAMERON *et al.* (eds.), *International Handbook of Educational Systems,* 3 vol. (1983–84); HAROLD E. MITZEL (ed.), *Encyclopedia of Educational Research,* 5th ed., 4 vol. (1982); TORSTEN HUSÉN and T. NEVILLE POSTLETHWAITE (eds.), *The International Encyclopedia of Education: Research and Studies,* 10 vol. (1985), with supplementary volumes, the first of which appeared in 1989; and GEORGE THOMAS KURIAN (ed.), *World Education Encyclopedia,* 3 vol. (1988).

Major trends and practical problems of education across the world are discussed in THOMAS F. GREEN, *The Activities of Teaching* (1971); GILBERT R. AUSTIN, *Early Childhood Education: An International Perspective* (1976); ISABELLE DEBLÉ, *The School Education of Girls: An International Comparative Study on School Wastage Among Girls and Boys at the First and Second Levels of Education* (1980); DIETMAR ROTHERMUND and JOHN SIMON (eds.), *Education and the Integration of Ethnic Minorities* (1986); JAMES A. BANKS and JAMES LYNCH (eds.), *Multicultural Education in Western Societies* (1986); EDMUND J. KING, *Other Schools and Ours,* 5th ed. (1979); J.R. HOUGH (ed.), *Educational Policy: An International Survey* (1984); ROBERT F. LAWSON (ed.), *Changing Patterns of Secondary Education: An International Comparison* (1987); DANIEL C. LEVY (ed.), *Private Education: Studies in Choice and Public Policy* (1986); ALEXANDER N. CHARTERS *et al., Comparing Adult Education Worldwide* (1981); NELL P. EURICH, *Systems of Higher Education in Twelve Countries* (1981); BURTON R. CLARK, *The Higher Education System: Academic Organization in Cross-National Perspective* (1983); and PHILIP H. COOMBS, *The World Crisis in Education: The View from the Eighties* (1985).

Studies of various contemporary educational philosophies and trends include JOHN DEWEY, *Democracy and Education: An Introduction to the Philosophy of Education* (1916, reprinted 1966); HARRY S. BROUDY, *Building a Philosophy of Education,* 2nd ed. (1961, reprinted 1977), and *The Uses of Schooling* (1988); PAUL H. HIRST, *Knowledge and the Curriculum: A Collection of Philosophical Papers* (1974); MERLE CURTI, *The Social Ideas of American Educators* (1935, reprinted 1978); MADAN SARUP, *Marxism and Education* (1978); JONAS F. SOLTIS (ed.), *Philosophy and Education* (1981); and ERNEST STABLER, *Founders: Innovators in Education, 1830–1980* (1986).

Works on individual countries are legion, and only a sample can be cited here. For Europe, see KEITH EVANS, *The Development and Structure of the English School System* (1985); and BRIAN SIMON and WILLIAM TAYLOR, *Education in the Eighties: The Central Issues* (1981), focusing on Great Britain; CHRISTOPH FÜHR, *Education and Teaching in the Federal Republic of Germany* (1979; originally published in German, 1979); W.D. HALLS, *Education, Culture, and Politics in Modern France* (1976); and LEON BOUCHER, *Tradition and Change in Swedish Education* (1982).

Studies specifically on U.S. education include LAWRENCE A. CREMIN, *The Transformation of the School: Progressivism in American Education, 1876–1957* (1961); SAMUEL BOWLES and HERBERT GINTIS, *Schooling in Capitalist America* (1976); ERNEST L. BOYER, *High School: A Report on Secondary Education in America* (1983); CHRISTOPHER JENCKS *et al., Inequality: A Reassessment of the Effect of Family and Schooling in America* (1972); CLARENCE J. KARIER, PAUL C. VIOLAS, and JOEL SPRING, *Roots of Crisis: American Education in the Twentieth Century* (1972); MICHAEL B. KATZ, *Class, Bureaucracy, and Schools: The Illusion of Educational Change in America,* expanded ed. (1975); JUDY JOLLEY MOHRAZ, *The Separate Problem: Case Studies of Black Education in the North, 1900–1930* (1979); DIANE RAVITCH, *The Troubled Crusade: American Education, 1945–1980* (1983); and FRED F. HARCLEROAD and ALLAN W. OSTAR, *Colleges and Universities for Change: America's Comprehensive Public State Colleges and Universities* (1987). ALLAN BLOOM, *The Closing of the American Mind* (1987), provides an example of intellectual criticism of the educational system.

For Canada, see CAROLYN COSSAGE, *A Question of Privilege: Canada's Independent Schools* (1977); ROBIN S. HARRIS, *A History of Higher Education in Canada, 1663–1960* (1976); ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Reviews of National Policies for Education: Canada* (1976); HUGH A. STEVENSON and J. DONALD WILSON, *Quality in Canadian Public Education: A Critical Assessment* (1988); T.H.B. SYMONS, *To Know Ourselves: The Report of the Commission on Canadian Studies,* 3 vol. in 2 (1975–84); and GEORGE S. TOMKINS, *A Common Countenance: Stability and Change in the Canadian Curriculum* (1986). For Australia, see PETER DWYER, BRUCE WILSON, and ROGER WOOK, *Confronting School and Work: Youth and Class Cultures in Australia* (1984); L.E. FOSTER, *Australian Education: A Sociological Perspective* (1981); PETER KARMEL (ed.), *Education, Change, and Society* (1981), papers of a conference of the Australian Council for Educational Research; and R.J.R. KING and R.E. YOUNG, *A Systematic Sociology of Australian Education* (1986). For New Zealand, see IAN CUMMING and ALAN CUMMING, *History of State Education in New Zealand, 1840–1975* (1978); and ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Reviews of National Policies for Education: New Zealand* (1983).

There are many works discussing the systems of education in those countries that have experienced major social upheavals. For the Soviet Union, see JOSEPH I. ZAJDA, *Education in the USSR* (1980); SHEILA FITZPATRICK, *Education and Social Mobility in the Soviet Union, 1921–1934* (1979); LUDWIG LIEGLE, *The Family's Role in Soviet Education* (1975; originally published in German, 1970); MERVYN MATTHEWS, *Education in the Soviet Union: Policies and Institutions Since Stalin* (1982); JOHN DUNSTAN, *Paths to Excellence and the Soviet School* (1978); and J.J. TOMIAK (ed.), *Soviet Education in the 1980s* (1983). For China, see THEODORE E. HSIAO, *The History of Modern Education in China* (1932); RONALD F. PRICE, *Education in Modern China,* 2nd ed. (1979); THEODORE HSI-EN CHEN, *Chinese Education Since 1949* (1981), *The Maoist Educational Revolution* (1974), and "Educational Development in the People's Republic of China, 1949–1981," in HUNGDAH CHIU and SHAO-CHUAN LENG (eds.), *China Seventy Years After the 1911 Hsin-Hai Revolution* (1984), pp. 364–389; WOLFGANG FRANKE, *The Reform and Abolition of the Traditional Chinese Examination System* (1960, reprinted 1972); KNIGHT BIGGERSTAFF, *The Earliest Modern Government School in China* (1961, reprinted 1972); and RUTH HAYHOE, *China's Universities and the Open Door* (1988). RONALD F. PRICE, *Marx and Education in Russia and China* (1977), is a comparative philosophical study.

Afro-Asian patterns of education are studied in ROBERT LEESTMA *et al., Japanese Education Today: A Report from the U.S. Study of Education in Japan* (1987); JAPAN PROVISIONAL COUNCIL ON EDUCATIONAL REFORM, *First Report on Educational Reform* (1985); RICHARD LYNN, *Educational Achievement in Japan: Lessons for the West* (1988); R. MURRAY THOMAS and T. NEVILLE POSTLETHWAITE (eds.), *Schooling in East Asia: Forces of Change: Formal and Nonformal Education in Japan, the Republic of China, the People's Republic of China, South Korea, North Korea, Hong Kong, and Macau* (1983), *Schooling in the ASEAN Region: Primary and Secondary Education in Indonesia, Malaysia, the Philippines, Singapore, and Thailand* (1980), and *Schooling in the Pacific Islands: Colonies in Transition* (1984); PAKISTAN. MINISTRY OF EDUCATION, *National Education Policy and Implementation Programme* (1979); ASIAN PROGRAMME OF EDUCATIONAL INNOVATION FOR DEVELOPMENT, *Towards Universalisation of Primary Education in Asia and the Pacific: Country Studies,* 12 vol. (1984), a UNESCO publication covering Bangladesh, China, India, Indonesia, Nepal, Pakistan, Papua New Guinea, the Philippines, South Korea, Vietnam, Sri Lanka, and Thailand; A. BISWAS and S.P. AGRAWAL (comps.), *Development of Education in India: A Historical Survey of Educational Documents Before and After Independence* (1986); S.N. MUKERJI, *Education in India Today and Tomorrow,* 7th ed. (1976); R.M. RUPERTI, *The Education System in Southern Africa* (1976; originally published in Afrikaans, 1974); PAM CHRISTIE, *The Right to Learn: The Struggle for Education in South Africa* (1985); and A.L. BEHR, *New Perspectives in South African Education* (1984).

Education in developing countries is the subject of A.R. THOMPSON, *Education and Development in Africa* (1981); A. BABS FAFUNWA and J.U. AISIKU (eds.), *Education in Africa: A Comparative Survey* (1982); DAVID G. SCANLON (ed.), *Church, State, and Education in Africa* (1966); ALI A. MAZRUI, *Political Values and the Educated Class in Africa* (1978); R.H. DAVE, A. OUANE, and A.M. RANAWEERA (eds.), *Learning Strategies for Post-Literacy and Continuing Education in Algeria, Egypt, and Kuwait* (1987); JUDITH COCHRAN, *Education in Egypt* (1986); JAMES ALLMAN, *Social Mobility, Education, and Development in Tunisia* (1979); JOSEPH S. SZYLIOWICZ, *Education and Modernization in the Middle East* (1973); BYRON G. MASSIALAS and SAMIR AHMED JARRAR, *Education in the Arab World* (1983); JOSEFINA VÁZQUEZ, *Nacionalismo y educación en México,* 2nd ed. (1975); GEORGE R. WAGGONER and BARBARA ASHTON WAGGONER, *Education in Central America* (1971); FAY HAUSSMAN and JERRY HAAR, *Education in Brazil* (1978); and DANIEL C. LEVY, *Higher Education and the State in Latin America* (1986).

(N.S./S.N.M./T.H.C./J.Bo./R.B./H.F.Gr./J.S.Sz./J.J.Ch./J.Z.V./R.F.L./O.A./Da.G.S./R.M.T.)

Egypt

Egypt (Arabic Miṣr), or the Arab Republic of Egypt (Jumhūrīyah Miṣr al-ʿArabīyah), as it has been known since 1971, has a total area of about 385,230 square miles (997,740 square kilometres). Its land frontiers border Libya in the west, The Sudan in the south, and Israel in the northeast. (Israeli forces occupied the Sinai Peninsula and the Gaza Strip in eastern Egypt after the Arab–Israeli War of 1967. In 1982 the Sinai was returned to Egypt.) In the north its Mediterranean coastline is about 620 miles (1,000 kilometres) and in the east its coastline on the Red Sea and the Gulf of Aqaba is about 1,200 miles. The capital is Cairo.

Egypt was the home of one of the principal civilizations of the ancient Near East and, like Mesopotamia, of one of the very earliest urban and literate societies. Its culture had an important influence on both ancient Israel and ancient Greece, which in turn helped to form the civilization of the modern West. Egypt also provided Africa with its earliest civilization and may well have had considerable influence on the development of other African cultures.

The special character evident in the civilization of ancient Egypt over a period of 3,000 years developed very rapidly at the time when the country first achieved unity. This great event happened in about 3100 BC, and while some of the seeds of Egyptian culture had sprouted before this time, it is proper to regard the start of the 1st dynasty as the virtual beginning of Egypt as the country and its civilization are now generally envisaged.

Perhaps the first and most important quality that typified this civilization was continuity. In every aspect of Egyptian life, in every manifestation of its culture, a deep conservatism can be observed. This clinging to the traditions and ways of earlier generations was the particular strength of the Egyptians. It can also be regarded as a weakness; but for a relatively primitive culture there was more to be gained than lost in attachment to the past. Regularity was a built-in characteristic of Egypt; life in the Nile Valley was determined to a great extent by the behaviour of the river itself. The pattern of inundation and falling water, of high Nile and low Nile, established the Egyptian year and controlled the lives of the Egyptian farmers—and most Egyptians were tied to a life on the land—from birth to death, from century to century. On the regular behaviour of the Nile rested the prosperity, the very continuity, of the land. The three seasons of the Egyptian year were even named after the land conditions produced by the river; *akhet,* the "inundation"; *peret,* the season when the land emerged from the flood; and *shomu,* the time when water was short. When the Nile behaved as expected, which most commonly was the case, life went on as normal; when the flood failed or was excessive, disaster followed.

Egypt has always been a hub for routes—westward along the coast of North Africa, northwest to Europe, northeast to the Levant, south along the Nile to Africa, and southeast to the Indian Ocean and the Far East. This natural advantage was enhanced in 1869 by the opening of the Suez Canal. The concern of the European powers to safeguard this link for strategic and commercial reasons is probably the most important single factor influencing the history of Egypt since the 19th century. The increasing presence of the United States and the Soviet Union in the Mediterranean since World War II has kept Egypt in the spotlight of world concern. It is not, however, simply in the context of the balance of power in the Mediterranean but also in Africa and in the Indian Ocean that Egypt's significance must be assessed. In addition, Egypt occupies a central position in the Arabic-speaking world. The country's geopolitical importance has increased during the 20th century as Arab nationalism has become a powerful and emotional political force in the Middle East and North Africa.

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief. The topography of Egypt is dominated by the Nile. For about 750 miles of its northward course through the country, the river cuts its way through bare desert, its narrow valley a sharply delineated strip of green, abundantly fecund in contrast to the desolation that surrounds it. From Lake Nasser, the river's entrance into southern Egypt, to Cairo in the north, the Nile is hemmed into its trenchlike valley by bordering cliffs, but at Cairo these disappear, and the river begins to fan out into its delta. As many as seven branches of the river once flowed through the Delta, but its waters are now concentrated in two, the Damietta Branch to the east and the Rosetta Branch to the west. Though totally flat apart from an occasional mound projecting through the alluvium, the Delta is far from featureless; it is crisscrossed by a maze of canals and drainage channels.

The Nile divides the desert plateau through which it flows into two unequal sections—the Western Desert (Arabic aṣ-Ṣaḥrāʾ al-Gharbīyah), between the river and the Libyan frontier; and the Eastern Desert (Arabic aṣ-Ṣaḥrāʾ ash-Sharqīyah), extending to the Suez Canal, the Gulf of Suez, and the Red Sea. Each of them has its own character, as does the third and smallest of the Egyptian deserts, the Sinai. The Western (Libyan) Desert is arid and without wadis (dry beds of seasonal rivers), while the Eastern Desert is extensively dissected by wadis and fringed by rugged mountains in the east. The desert of central Sinai is open country, broken by isolated hills and scored by wadis.

The mountains

Egypt is not, as is often believed, an unrelievedly flat country. Mountainous areas occur in the extreme southwest of the Western Desert, along the Red Sea coast, and in southern Sinai. The high ground in the southwest is associated with the ʿUwaynāt mountain mass, which lies just outside Egyptian territory. A number of peaks in the Red Sea Hills (Itbāy) rise to more than 6,000 feet (1,800 metres), and the highest, Mount Shāiyb al-Banāt, reaches 7,175 feet (2,187 metres). The sharply serrated crests of the mountains of southern Sinai reach elevations of more than 8,000 feet; among them is Mount Catherine (Jabal Kātrīnā), Egypt's highest mountain, which has an elevation of 8,668 feet (2,642 metres).

The coastal regions of Egypt, with the exception of the Delta, are everywhere hemmed in either by desert or by mountain; they are arid or of very limited fertility. The coastal plain, in both the north and east, tends to be narrow; it seldom exceeds a width of 30 miles. With the exception of the cities of Alexandria, Port Said, and Suez and a few small ports and resorts, the coastal regions are sparsely populated and underdeveloped.

Drainage and soils. Apart from the Nile, the only natural perennial surface drainage consists of a few small streams in the mountains of southern Sinai. Most of the valleys of the Eastern Desert drain westward to the Nile. They are eroded by water but normally dry; only after heavy rainstorms in the Red Sea Hills do they carry torrents. The shorter valleys on the eastern flank of the Red Sea Hills drain toward the Red Sea; they, too, are normally dry. Drainage in the Sinai mountains is toward the gulfs of Suez and Aqaba; as in the Red Sea Hills, torrent action has produced valleys that are deeply eroded and normally dry.

The central plateau of Sinai drains northward toward Wadi al-ʿArīsh, a depression in the desert that occasionally carries surface water. One of the features of the Western Desert is its aridity, as shown by the absence of drainage lines. There is, however, an extensive water table beneath the Western Desert. Where the water table comes near the surface it has been tapped by wells in some oases.

Outside the areas of Nile silt deposits, the nature of such cultivable soil as exists depends upon the availability of the water supply and the type of rock in the area. Almost one-third of the total land surface of Egypt consists of Nubian sandstone, which extends over the southern sections of both the Eastern and Western deserts. Limestone deposits of the Eocene Epoch (from 38,000,000 to 54,000,000 years old) cover a further one-fifth of the land surface, including central Sinai and the central portions of both the Eastern and Western deserts. The northern part of the Western Desert consists of Miocene limestone (from 7,000,000 to 26,000,000 years old). About one-eighth of the total area, notably the mountains of Sinai, the Red Sea, and the southwest part of the Western Desert, consists of ancient igneous and metamorphic rocks.

The Nile silt

The silt, which constitutes the present-day cultivated land in the Delta and the Nile Valley, has been carried down from the Ethiopian Highlands by the Nile's upper tributary system, consisting of the Blue Nile and the ʿAṭbarah rivers. The depth of the deposits ranges from more than 30 feet in the northern Delta to about 22 feet at Aswān. The White Nile, which is joined by the Blue Nile at Khartoum, in The Sudan, supplies important chemical constituents. The composition of the soil varies and is generally more sandy toward the edges of the cultivated area. A high clay content makes it difficult to work, and a concentration of sodium carbonate sometimes produces infertile black-alkali soils. In the north of the Delta, salinization has produced the sterile soils of the so-called *barārī* ("barren") regions.

Climate. Egypt lies within the North African desert belt; its general climatic characteristics, therefore, are low annual rainfall and a considerable seasonal and diurnal (daily) temperature range, with sunshine occurring throughout the year. In the desert, cyclones stir up sand or dust storms, called khamsins, which occur most frequently from March to June; these are caused by tropical air from the south that moves northward as a result of the extension northeastward of the low-pressure system of The Sudan. A khamsin is accompanied by a sharp increase in temperature of from 14° to 20° F (8° to 11° C), a drop in relative humidity (often to 10 percent), and thick dust; it can reach gale force.

The climate is basically biseasonal, with winter lasting from November to March and summer from May to September, with short transitional periods intervening. The winters are cool and mild, and the summers are hot. Mean January minimum and maximum temperatures show a variation of between 48° and 65° F (9° and 18° C) in Alexandria and 48° and 74° F (9° and 23° C) at Aswān. The summer months are hot throughout the country, with mean midday June maximum temperatures ranging from 91° F (33° C) at Cairo to 106° F (41° C) at Aswān. Egypt enjoys a very sunny climate, with some 12 hours of sunshine per day in the summer months and between eight and 10 hours per day in winter. Extremes of temperature can occur, and prolonged winter cold spells or summer heat waves are not uncommon.

Humidity diminishes noticeably from north to south and on the desert fringes. Along the Mediterranean coast the humidity is high throughout the year, but it is highest in summer. When high humidity levels coincide with high temperatures, oppressive conditions result.

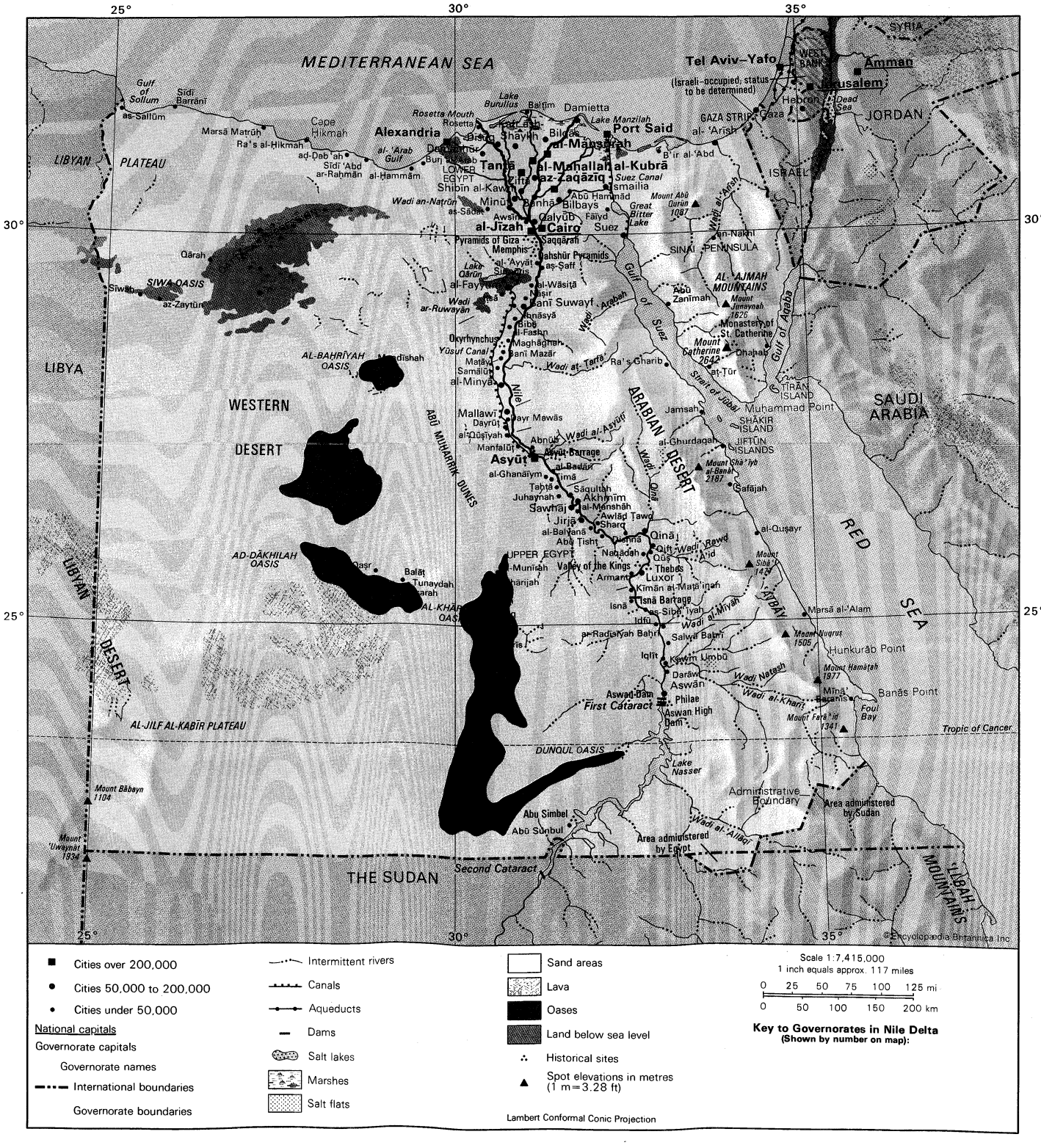
The rainfall in Egypt occurs largely in the winter months; it is meagre on average but highly variable. The amount diminishes sharply southward; the annual average at Alexandria is about seven inches (178 millimetres), Cairo

has about one inch, and Aswān receives only about one-tenth of an inch. The Red Sea coastal plain and the Western Desert are almost rainless. The Sinai Peninsula receives somewhat more rainfall: the northern sector has an annual average of about five inches.

Plant and animal life. In spite of the lack of rainfall, the natural vegetation of Egypt is varied. Much of the Western Desert is totally devoid of plant life of any kind, but where some form of water exists the usual desert growth of perennials and grasses is found; the coastal strip has a rich plant life in spring. The Eastern Desert receives sparse rainfall; it supports a varied vegetation that includes tamarisk, acacia, and *markh* (a leafless, thornless tree with bare branches and slender twigs), as well as a great variety of thorny shrubs, small succulents, and aromatic herbs. This growth is even more striking in the wadis of the Red Sea Hills and of Sinai and in the Elba Mountains in the southeast.

Water plants

The Nile and irrigation canals and ditches support many varieties of water plants; the lotus of antiquity is to be found in drainage channels in the Delta. There are more than 100 kinds of grasses, among them bamboo and *ḥalfā'* (a coarse, long grass growing near water). Robust perennial reeds such as the Spanish reed and the common reed are widely distributed in Lower Egypt, but the papyrus, cultivated in antiquity, is now found only in botanical gardens.



MAP INDEX

The date palm, both cultivated and subspontaneous, is found throughout the Delta, in the Nile Valley, and in the oases. The doum palm (an African fan palm) is identified particularly with Upper Egypt and the oases, although there are scattered examples elsewhere.

There are very few native trees. The Phoenician juniper is the only native conifer, although there are several cultivated conifer species. The acacia is widely distributed, as are eucalyptus and sycamore. The casuarina, one of the most important timber trees in the country, was introduced in the 19th century. Other foreign importations, such as jacaranda, poinciana (a tree with orange or scarlet flowers), and lebbek (a leguminous tree), have become a characteristic feature of the Egyptian landscape.

Domestic animals include buffalo, camels, donkeys, sheep, and goats, the last of which are particularly noticeable in the Egyptian countryside. The animals that figure so prominently on the ancient Egyptian friezes—hippopotamuses, giraffes, and ostriches—no longer exist in Egypt; crocodiles are found only south of the Aswān High Dam. The largest wild animal is the mountain sheep, which survives in the southern fastnesses of the Western Desert. Other desert animals are the dorcas gazelle, the miniature desert fox, the mountain goat, the Egyptian hare, and two kinds of jerboa (a mouselike rodent with long hindlegs for jumping). The Egyptian jackal still exists, and the cony (a small rodent) is found in the Sinai mountains. There are two carnivorous mammals: a species of wildcat and the striped Egyptian mongoose. Several varieties of lizard are found, including the large monitor. Poisonous snakes include more than one species of viper; the speckled snake is found throughout the Nile Valley and the Egyptian cobra in agricultural areas. Scorpions are common in desert regions. There are numerous species of rodents, among which can be found the powerfully built Pharaoh's rat. Many varieties of insects are to be found, including the Egyptian locust.

Bird life

Egypt is rich in bird life. Many birds pass through in large numbers on their spring and autumn migrations; in all, there are more than 200 migrating types to be seen, as well as more than 150 resident birds. The hooded crow is a familiar resident, and the black kite is a characteristic resident along the Nile Valley and in al-Fayyūm. Among the birds of prey are the lanner falcon and the kestrel. Lammergeier and golden eagles are residents of the Eastern Desert and Sinai. The sacred ibis (a long-billed wading bird) is no longer found, but the great egret and buff-backed heron are residents of the Nile Valley and al-Fayyūm, as is the hoopoe (a bird with an erectile, fanlike crest). Resident desert birds are a distinct category, numbering about 24 kinds.

The Nile contains about 190 varieties of fish, the most common being *bulṭī* (a coarse-scaled, spiny-finned fish) and the Nile perch. The lakes on the Delta coast contain mainly *būrī* (gray mullet). Lake Qārūn in al-Fayyūm *muḥāfaẓah* (governorate) has been stocked with *būrī,* and Lake Nasser with *bulṭī,* which grow very large in its waters.

Settlement patterns. Physiographically, Egypt is usually divided into four major regions—the Nile Valley and Delta, the Eastern Desert, the Western Desert, and Sinai. When both physical and cultural characteristics are considered together, however, the country may be divided into six subregions—the Nile Delta; the Nile Valley from Cairo to south of Aswān; the Nubian Valley (since the early 1970s filled by Lake Nasser); the Eastern Desert and the Red Sea coast; Sinai; and the Western Desert and its oases.

The Delta. The Nile Delta, or Lower Egypt, covers an area of 9,650 square miles. It is 100 miles long from Cairo to the Mediterranean, with a coastline stretching 150 miles from Alexandria to Port Said. Much of the Delta coast is taken up by the brackish lagoons of Lakes Maryūṭ, Idkū, Burullus, and Manzilah. The conversion of the Delta to perennial irrigation has made possible the raising of two or three crops a year, instead of one, over more than half of its total area.

The Delta population

About half of the population of the Delta are peasants (fellahin)—either small landowners or labourers—living on the produce of the land. The remainder live in towns or cities, the largest of which is Cairo. As a whole, they have had greater contact with the outside world, particularly with the rest of the Middle East and Europe, than the inhabitants of the more remote southern valley and are generally less traditional and conservative.

The Valley. The cultivated portion of the Nile Valley between Cairo and Aswān varies from five to 10 miles in width, although there are places where it narrows to a few hundred yards and others where it broadens to 14 miles. Since the completion of the Aswān High Dam in 1970, the 2,500,000-acre valley has been under perennial irrigation. The inhabitants of the Valley from Cairo up to Aswān *muḥāfaẓah* are referred to as Ṣaʿīdī (Upper Egyptians) and are more conservative than the Delta people. In some areas women still do not appear in public without a veil; family honour is very important, and vendetta laws apply. Until the building of the High Dam, the Aswān *muḥāfaẓah* was one of the poorest in the Valley and the most remote from outside influences.

The Nubian Valley, or Lake Nasser. Until it was flooded by the waters impounded behind the High Dam to form Lake Nasser, the Nubian Valley of the Nile extended for 160 miles between the town of Aswān and the Sudanese border—a narrow and picturesque gorge with a limited cultivable area. The 100,000 inhabitants were resettled, mainly in the government-built villages of New Nubia, at Kawm Umbū (Kom Ombo), north of Aswān. Lake Nasser was developed during the 1970s for its fishing and as a tourist area, and settlements have grown up around it.

The Eastern Desert. The Eastern Desert comprises almost one-fourth of the land surface of Egypt and covers an area of about 85,690 square miles. The northern tier is a limestone plateau, consisting of rolling hills, stretching from the Mediterranean coastal plain to a point roughly opposite Qinā on the Nile. Near Qinā, the plateau breaks up into cliffs about 1,600 feet high and is deeply scored by wadis, which make the terrain very difficult to traverse. The outlets of some of the main wadis form deep bays, which contain small settlements of seminomads. The second tier includes the sandstone plateau from Qinā southward. The plateau is also deeply indented by ravines, but they are relatively free from obstacles, and some are usable as routes. The third tier consists of the Red Sea Hills and the Red Sea coastal plain. The hills run from near Suez to the Sudanese border; they are not a continuous range but consist of a series of interlocking systems more or less in alignment. They are geologically complex, with ancient igneous and metamorphic rocks. These include granite that, in the neighbourhood of Aswān, extends across the Nile Valley to form the First Cataract—that is, the first set of rapids on the river. At the foot of the Red Sea Hills the narrow coastal plain widens southward, and parallel to the shore there are almost continuous coral reefs. In popular conception and usage, the Red Sea Littoral can be regarded as a subregion in itself.

The majority of the sedentary population of the Eastern Desert live in the few towns and settlements along the coast, the largest being Ra's Gharib. No accurate figures are available for the nomadic population, but they are believed to constitute about 12 percent of the region's total population. They belong to various tribal groups, the most important being—from north to south—the Ḥuwayṭāt, Maʿāzah, ʿAbābdah, and Bishārīn. There are more true nomads in the Eastern than the Western Desert because of the greater availability of pasture and water. They live either by herding goats, sheep, and camels or by trading—mainly with mining and petroleum camps or with the fishing communities on the coast.

Nomads of the Eastern Desert

The Western Desert. The Western Desert comprises two-thirds of the land surface of Egypt and covers an area of about 262,800 square miles. From its highest altitude—more than 3,300 feet—on the plateau of al-Jilf al-Kabīr in the southeast, the rocky plateau slopes gradually northeastward to the first of the depressions that are a characteristic feature of the Western Desert—that containing the oases of al-Khārijah and ad-Dākhilah. Farther north are the hollows containing the oases of al-Farāfirah and al-Baḥrīyah. Northwestward from the latter the plateau continues to fall toward the Qattara Depression (Munkhafaḍ al-Qaṭṭārah), which is uninhabited. West of the Qattara Depression and near the Libyan border is the largest and most populous oasis, that of Siwa. It has been inhabited for thousands of years and is less influenced by modern development. South of the Qattara Depression, and extending west to the Libyan border, the Western Desert is composed of great ridges of blown sand, interspersed with stony tracts. Beyond the Qattara Depression northward, the edge of the plateau follows the Mediterranean, leaving a narrow coastal plain.

Outside the oases, the habitable areas of the Western Desert, mainly near the coast, are occupied by the Awlād ʿAlī tribe. Apart from small groups of camel herders in the south, the population is no longer totally nomadic. Somewhat less than half are seminomadic herdsmen; the remainder are settled and, in addition to maintaining herds of sheep and goats, pursue such activities as fruit growing, fishing, trading, and handicrafts.

The Western Desert supports a much larger population than the Eastern Desert. Maṭrūḥ, an important summer resort on the Mediterranean, is the only urban centre. Other scattered communities are found mainly near railway stations and along the northern cultivated strip.

The oases, though geographically a part of the Western Desert, are ethnically and culturally distinct. The southern oases of al-Khārijah and ad-Dākhilah have been developed to some extent as part of a reclamation project centred on exploiting underground water resources. Other oases are al-Farāfirah, al-Baḥrīyah, and Siwa.

Sinai. Sinai comprises a wedge-shaped block of territory with its base along the Mediterranean coast and its apex bounded by the Gulfs of Suez and Aqaba; it covers an area of approximately 23,000 square miles. Its southern portion consists of rugged, sharply serrated mountains. The central area of Sinai consists of two plateaus, at-Tīh and al-ʿAjmah, both deeply indented and dipping northward toward Wadi al-ʿArīsh. Toward the Mediterranean, the northward plateau slope is broken by dome-shaped hills; between them and the coast are long, parallel lines of dunes, some of which are more than 300 feet high. The most striking feature of the coast itself is the 60-mile-long salt lagoon, Lake Bardawīl.

The majority of the population are Arabs, many of whom have settled around al-ʿArīsh and in the northern coastal area, although substantial numbers in the central plateau and the Sinai mountains continue to be nomadic or seminomadic. Another concentration of sedentary population is found at al-Qanṭarah, on the east side of the Suez Canal.

Rural settlement. The settled Egyptian countryside, throughout the Delta and the Nile Valley to the High Dam, exhibits great homogeneity, although minor variations occur from north to south.

Villages of the Delta and the Nile Valley

The typical rural settlement is a compact village surrounded by intensively cultivated fields. The villages range in population from 500 to more than 10,000. They are basically similar in physical appearance and design, except for minor local variations in building materials, design, and decoration. The date palm, sycamore, eucalyptus, and casuarina are common features of the landscape. Until comparatively recently, the only source of drinking water was the Nile; in consequence, many of the villages are built along the banks of its canals. Some of the oldest villages are situated on mounds—a relic of the days of basin irrigation and annual flooding.

In the Delta the houses, one or two stories high, are built of mud bricks plastered with mud and straw; in the southern parts of the Valley more stone is used. The houses are joined to one another in a continuous row. In a typical house the windows consist of a few small round or square openings, barely permitting enough air or light to enter. The roofs are flat, built of layers of dried date leaves, with date-palm rafters; they are used to store corn (maize) and cotton stalks, as well as dung cakes used for fuel. Roofs are also a favourite sleeping place on hot summer nights. For grain storage small cone-shaped silos of plastered mud are built on the roof and are then sealed to prevent the ravages of insects and rodents.

The houses of the poorer peasants usually consist of a narrow passageway, a bedroom, and a courtyard; part of the courtyard may be used as an enclosure for farm animals. Furniture is sparse. Ovens are made of plastered mud and are built into the wall of the courtyard or inside the house. In the larger and more prosperous villages, houses are built of burnt bricks reinforced with concrete, are more spacious, and often house members of an extended family. Furniture, running water, bathroom installations, and electricity are additional signs of prosperity.

Typical features of the smaller Egyptian village, in both the Delta and the Valley, are the mosque or the church, the primary school, the decorated pigeon cote, service buildings belonging to the government, and a few shops. Most of the people in the smaller villages are engaged in agriculture. In the larger villages, there may be some professional and semiprofessional inhabitants as well as more artisans, skilled workers, and shopkeepers. Outside the larger settlements, "combined service units"—consisting of modern buildings enclosing the social service unit, village cooperative, health unit, and school—are sometimes found, standing in striking contrast to the mud houses of the village itself.

The rural life-style

The population density of the inhabited area is such that the presence of people is obvious everywhere, even in the open countryside. In the early morning and the late afternoon, the peasants can be seen in large numbers on the roads, going to or coming from the fields with their farm animals. During the entire day the men, with their long tunics (*gallābīyahs*) tucked up around the waist, can be seen working the land with age-old implements such as the *fās* (hoe) and *minjal* (sickle); occasionally a modern tractor is seen. In the Delta older women in long, black robes, younger ones in more colourful cottons, and children over six years of age help with the less laborious tasks. In some parts of the Valley, however, women over age 16 do not work in the field, and their activities are confined to the household. They seldom appear in public except with a black muslin headdress covering their heads and faces. Young children can be seen everywhere—an omnipresent reminder of the high birthrate.

Unless situated on a highway, villages are reached by unpaved dirt roads. Inside the villages the roads consist mainly of narrow, winding footpaths. All villages, however, have at least one motorable road.

The Western Desert oases are not compact villages but small, dispersed agglomerations surrounded by green patches of cultivation; they are often separated from each other by areas of sand. Al-Khārijah, for example, is the largest of five scattered villages. Traditionally, the houses in the oases were up to six stories high, made of packed mud, and clustered close together for defense. Modern houses are usually two stories high and farther apart.

Urban settlement. Although for census purposes Egyptian towns are considered to be urban centres, some of them are overgrown villages, containing large numbers of peasants and persons engaged in work relating to agriculture and rural enterprises. Some of the towns that have acquired urban status in the second half of the 20th century continue to be largely rural, although they have government officials, people engaged in trade and commerce, industrial workers, technicians, and professional people among their residents. One characteristic of towns and, indeed, of the larger cities is their rural fringe. Towns and cities have grown at the expense of agricultural land, with urban dwellings and apartment buildings mushrooming haphazardly among the fields. There is little evidence of town or city planning or of adherence to building regulations; often mud village houses are embraced within the confines of a city.

Urban characteristics

Al-Qaṣr, in the oasis of ad-Dākhilah in the Western Desert.
© Georg Gerster—Photo Researchers, Inc.

Buildings in towns and smaller cities are usually two-storied houses or apartment blocks four to six stories high. The better ones are lime washed, with flat roofs and numerous balconies; other houses and buildings are often of unpainted red brick and concrete.

Whereas most of the cities of Egypt do not have many distinctive features, some such as Cairo, Alexandria, and Aswān have special characteristics of their own. Cairo is a complex and crowded metropolis, with architecture representing more than 1,000 years of history. Greater Cairo (including al-Jīzah and other suburban settlements) and Alexandria, together with the most important towns along the Suez Canal—Port Said, Ismailia, and Suez—are modern and Western in appearance. Extensive rebuilding of the towns in the canal zone, severely damaged in the fighting between 1967 and 1973, followed the peace treaty with Israel in 1979.

THE PEOPLE

The use of Arabic

Linguistic composition. For almost 13 centuries Arabic has been the written and spoken language of Egypt. Before the Arab invasion in AD 639, Coptic, the language descended from ancient Egyptian, was the language of both religious and everyday life for the mass of the population; by the 12th century, however, it had been totally replaced by Arabic, continuing only as a liturgical language for the Coptic Orthodox Church. Arabic has become the language of both the Egyptian Christian and Muslim.

The written form of the Arabic language, in grammar and syntax, has remained substantially unchanged since the 7th century. In other ways, however, the written language has changed—the modern forms of style, word sequence, and phraseology are simpler and more flexible than in classical Arabic and are often directly derivative of English or French.

This modern literary Arabic, which is developing out of classical or medieval Arabic, is the lingua franca shared by educated persons throughout the Arab world. Alongside it there exist the various regional dialects of Arabic, which differ widely from it as well as from one another. Within the amorphous grouping referred to as Egyptian colloquial, a number of separate dialects can be discerned—each fairly homogeneous but with further strata of variation within the group. One of these is the dialect of the Bedouin of the Eastern Desert and of Sinai; the Bedouin of the Western Desert constitute a separate dialect group. Upper Egypt has its own vernacular, markedly different from that of Cairo. The Cairo dialect is used, with variations, throughout the towns of the Delta; the rural people have their own vernacular. Direct contact with foreigners over a long period has led to the incorporation of many loanwords into Cairene colloquial Arabic. The long contact with foreigners and the existence of foreign-language schools also explains the polyglot character of Egyptian society. Most educated Egyptians are fluent in English or French or both, in addition to Arabic.

Non-Arabic languages

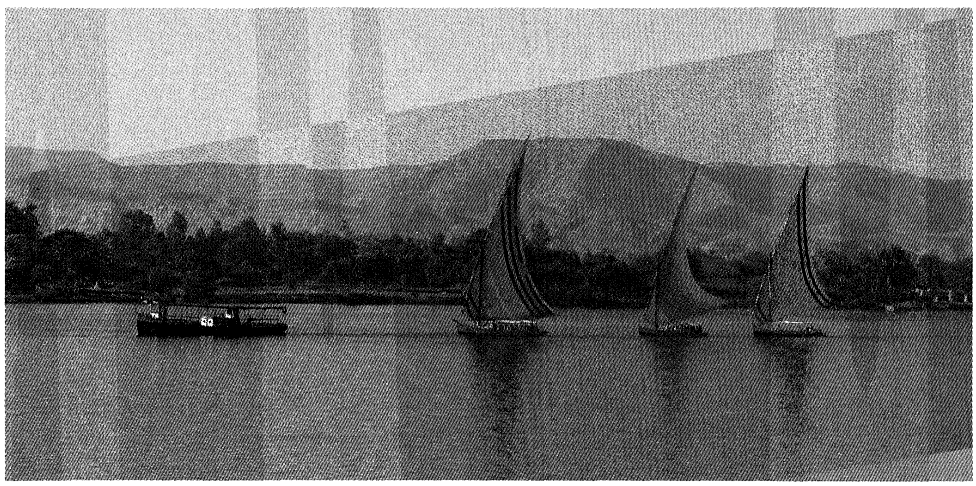
There are other minor linguistic groups. The Hamitic Beja (Bujah) of the southern section of the Eastern Desert use To Badawī. At Siwa Oasis in the Western Desert there are groups whose language is related to Berber. The Nubian speak a language containing both Sudanic and Hamitic features. There are other minority linguistic groups, notably Greek, Italian, and Armenian, although they are much smaller than they once were.

Ethnic composition. The population of the Nile Valley and the Delta (comprising about 99 percent of Egypt) forms a fairly homogeneous group whose dominant physical characteristics are the result of the admixture of the indigenous pre-Islāmic Hamitic–Armenoid population with Arab stock. The peasant, or fellah, is less racially mixed than the town dweller. In the towns—the northern Delta towns especially—the foreign invader, Persian, Roman, Greek, crusader, and Turk, has left behind a more heterogeneous mixture. The inhabitants of the middle Nile Valley up to Aswān, the Ṣaʿīdī (Upper Egyptians), are of the same racial stock as the inhabitants of the Delta but have darker skin and are slightly taller and have a sturdier build. Settled communities in Aswān *muḥāfaẓah* tend to be a mixture of Ṣaʿīdī, long-settled nomads such as the Jaʿāfirah and ʿAbābdah, and Nubian.

The Nubian, though having Arab blood, have preserved racial characteristics that are non-Arab. They are tall and thin, with Caucasoid features, and are of a much darker colouring than the average Egyptian. They also differ in that their kinship structure goes beyond the lineage; they are divided into clans and broader segments, whereas among other Egyptians of the Valley and Lower Egypt known members of the lineage are the only ones recognized as kin.

The deserts of Egypt contain nomadic, seminomadic, or sedentary but formerly nomadic groups, with distinct ethnic characteristics. Apart from a few tribal groups of non-Arab stock and the mixed urban population, the inhabitants of Sinai and the northern section of the Eastern Desert are all fairly recent immigrants from Arabia. Like the Arabian Bedouin, they are usually slightly built and brown skinned and have prominent, hooked noses. Their social organization is tribal, each group conceiving of itself as being united by a bond of blood and as having descended from a common ancestor. Originally tent dwellers and nomadic herders, many have become seminomads or even totally sedentary, as in northern Sinai.

The southern section of the Eastern Desert is inhabited by the Hamitic Beja. Though claiming Arab descent, they are of different racial stock, with oval faces, straight noses, and large eyes; they bear a distinct resemblance to the sur-



Feluccas on the Nile River, near Luxor in Upper Egypt.
© Robert Frerck—Odyssey Productions

viving depictions of predynastic Egyptians. The Egyptian Beja are divided into two tribes—the ʿAbābdah and the Bishārīn. The ʿAbābdah occupy the Eastern Desert south of a line between Qinā and al-Ghurdaqah; there are also several groups settled along the Nile between Aswān and Qinā. The Bishārīn live mainly in The Sudan, although some dwell in the Elba Mountain region, their traditional place of origin. Both the ʿAbābdah and Bishārīn people are nomadic pastoralists who tend herds of camels, goats, and sheep.

The Saʿādī and the Mūrābiṭīn

The inhabitants of the Western Desert, outside the oases, claim Arab descent but are a mixture of Arab and Berber stock. They are divided into two groups, the Saʿādī and the Mūrābiṭīn. The Saʿādī regard themselves as descended from Banū Hilāl and Banū Sulaymān, the great Arab tribes that immigrated into North Africa in the 11th century. The most important and numerous of the Saʿādī group are the Awlād ʿAlī. The Mūrābiṭīn clans occupy a client status in relation to the Saʿādī and may be descendants of the original Berber inhabitants of the region. Originally herders and tent dwellers, the Bedouin of the Western Desert have become either seminomadic or totally sedentary. They are not localized by clan, and members of a single group may be widely dispersed.

The original inhabitants of the oases of the Western Desert were of Berber stock. There has, however, been a considerable admixture of blood—Egyptian from the Nile Valley, Arab, Sudanese, Turkish, and, particularly in the case of al-Khārijah, black African—for this was the point of entry into Egypt of the caravan route from Darfur, the Darb al-Arbaʿīn.

In addition to the indigenous groups, there are in Egypt a number of small foreign ethnic groups. In the 19th century there was rapid growth of communities of unassimilated foreigners, mainly European, living in Egypt; these acquired a dominating influence over finance, industry, and government. In the 1920s, which was a peak period, the number of foreigners in Egypt was in excess of 200,000, the largest community being the Greeks, followed by the Italians, British, and French. Since Egypt's independence the size of the foreign communities has been greatly reduced.

Religions. Islām is the official religion of Egypt, and a large majority of the population embrace the Sunnī, or orthodox, branch of Islām. A strong sense of piety is a characteristic of the Egyptian Muslim. Prayer is observed punctiliously, particularly public prayer in the mosques, and fasting during the month of Ramaḍān (the ninth month of the Islāmic calendar) is strictly observed. Almsgiving and pilgrimage to Mecca are, if possible, also enjoined.

The Copts

The majority of the Christian population of Egypt are Copts. In language, dress, and way of life they are indistinguishable from Muslim Egyptians; their church ritual and traditions, however, date from before the Arab conquest in the 7th century. Ever since it broke with the Eastern Church in the 5th century, the Coptic Orthodox Church has maintained its autonomy, and its beliefs and ritual have remained basically unchanged. The Copts have traditionally been associated with certain handicrafts and trades and, above all, with accountancy, banking, commerce, and the civil service; there are, however, rural communities that are wholly Coptic, as well as mixed Coptic–Muslim villages. As a result of marrying almost exclusively within their community, many Copts preserve in their facial and body features the characteristics of the people of Pharaonic Egypt.

The Copts are most numerous in the middle Nile Valley *muḥāfaẓāt* of Asyūṭ, al-Minyā, and Qinā. About one-fourth of the total Coptic population lives in Cairo.

Among other religious groups are the Coptic Catholic, Greek Orthodox, Greek Catholic, Armenian Orthodox and Catholic, Maronite, Syrian Catholic, Anglican, and Protestant. There is also a small Jewish community.

Demographic trends. Most of Egypt's people live along the banks of the Nile River, where the population density, estimated to be more than 2,700 persons per square mile (1,100 persons per square kilometre), is one of the highest in the world. The rapidly growing population is young, with two-fifths of the total under 15 years of age. Despite improvements in health care, infant mortality is very high and about half of all deaths occur among children less than five years of age. Life expectancy, however, increased from only about 33 years in 1927 to almost 60 years by the mid-1980s. Almost half of the population lives in urban areas. (L.S.El H./Ma.J./D.H./C.G.S.)

THE ECONOMY

The economy of Egypt, according to the constitution of 1971, is one based on socialism, with the people controlling all means of production. The progress of socialism after 1952 was initially hesitant, despite land-reform measures, but it gathered momentum after 1961, when major nationalization steps were taken in an attempt to curb the private sector and destroy the political power of Egyptian capitalists. Until the early 1970s almost all important sectors of the economy either were public or were strictly controlled by the government. This included large-scale industry, communications, banking and finance, the cotton trade, foreign trade as a whole, and many other sectors. Private enterprise came gradually to find its scope restricted, but some room for maneuver was still left in real estate and in agriculture and, later, in the export trade. Personal income, as well as land ownership, was strictly limited by the government. Some of these restrictions have been relaxed, permitting greater private sector participation in various economic areas.

The public sector and the role of government. As the role of the private sector lessened in the 1950s and '60s, that of the government continuously expanded. The government, when not actually in possession, regulates all important aspects of production and distribution. It im-

poses controls on agricultural prices, controls rent, runs the internal trade, regulates foreign travel and the use of foreign exchange, and appoints and supervises the boards of directors of corporations. The government initiates projects and allocates investment. Although the everyday running of corporations is left to the boards of directors, these receive instructions from public boards, and the chairmen of boards receive their instructions from the appropriate minister. The government formulates five-year development plans to guide economic development.

Direct taxation

Taxation. With the majority of the population earning very low incomes, direct taxation falls on the few rich; income-tax rates are made sharply progressive in an attempt to achieve a degree of equality in income distribution. Direct taxes on income, mostly levied on businesses, account for about two-thirds of governmental revenue.

Trade unions and employer associations. Trade unions are closely controlled by the government. Workers obtain a share of the profits earned by corporations and elect their representatives to boards of directors; they are also heavily represented in the National Assembly. In all these activities, however, official selection works side by side with free elections. Trade unions are often vocally active in national policies but are seldom the instrument for negotiating higher wages or better work conditions. There are a few employers' associations, but they have little industrial power.

Contemporary economic policies. In the early 1970s the Egyptian government campaigned for increased foreign investment and began receiving financial aid from the oil-rich Arab states. Although Arab aid was suspended after the signing of the 1979 peace treaty with Israel, the subsequent return of several Western and Japanese corporations, associated with the normalization of Egyptian relations with Israel, increased the potential for further foreign investment in the country. Much of the effort exerted by the government in the early 1980s was devoted to adjusting the economy to the situation resulting from the 1979 Egyptian-Israeli peace treaty. With decreased expenditure on defense, increased allocations were made available for development. Egypt's economy began to be more resilient, primarily because of discoveries of oil and increased Western aid.

Increases in population have put pressure on resources, however, and underemployment has become endemic.

Living standards

Wages and cost of living. The general standard of living in Egypt is rather low; in relation to the size of its population, its economic resources are limited. Land remains its main source of natural wealth, but the amount of land is insufficient to support the population adequately. The realization of the need to curb the rate of population increase led, in 1964, to a national family planning program, which has had only limited success.

The rural population, especially the landless agricultural labourers, has the lowest standard of living in the country. Industrial and urban workers enjoy, on the whole, a higher standard. The highest wages are earned in such industries as the petroleum and manufacturing industries; many workers in industry receive additional benefits by way of social insurance and extra health and housing facilities. The salaries of professional groups are also low. Low wage levels have to some extent been offset by the low cost of living, but by the late 1970s this advantage was eliminated by high inflation rates.

Resources. About 96 percent of Egypt's total area is desert. Lack of forests, permanent meadows, or pastures places a heavy burden on the available arable land, which constitutes only about 3 percent of the total area. This limited area, which sustains on the average almost seven persons per acre, is, however, highly fertile and is cropped more than once a year. Although a large percentage of the population derives its livelihood from agriculture, a growing proportion of the labour force is engaged in manufacturing, and the contribution of the manufacturing and mining sectors to the domestic product has grown to twice that of agriculture—with service activities contributing most of the remainder. Because of the shortage of land, underemployment of labour began to be manifest in agriculture early in the 20th century, and the development of nonagricultural production since then has failed to provide full employment to the increasing labour force.

Mineral resources. Compared with the physical size of the country and the level of its population, the mineral resources of Egypt are scanty. The search for petroleum began earlier in Egypt than elsewhere in the Middle East, and production on a small scale began as early as 1908, but it was not until the mid-1970s that significant results were achieved. By the early 1980s Egypt had become an important oil producer, although total production was relatively small by Middle Eastern standards. Several of Egypt's major known phosphate deposits are mined at Isnā, Ḥamrāwayn, and Safājah. Coal deposits are located in the partially developed Maghara mines in the Sinai Peninsula. Manganese deposits located in the Eastern Desert have been the primary source for manganese production since 1967, and there are also reserves of manganese on the Sinai Peninsula. Egypt mines iron ore from deposits at Aswān, and development work has continued at al-Wāḥāt al-Baḥrīyah Oasis. Chromium, uranium, and gold deposits are also found in the country.

Biological resources. Egypt's biological resources, centred around the Nile, have long been one of its principal assets. There are no forests or any permanent vegetation of economic significance, apart from the land under cultivation. Water buffalo, cattle, asses, goats, sheep, and camels are the most important livestock. Animal husbandry and poultry production have continued to increase.

Hydroelectric and other power resources. The Nile constitutes an incomparable source of energy; further sources are represented by coal, oil, and gas deposits. Almost half of Egypt's electrical energy comes from thermal stations; hydroelectric plants, including those at the Aswān High Dam, supply the remainder.

Agriculture and fishing. Agriculture is an important sector of the Egyptian economy. It contributes substantially to the gross national product, employs a large part of the labour force, and provides the country—through agricultural exports—with an important part of its foreign exchange. Increased pressure of population has led to an intensification of cultivation almost without parallel elsewhere. Heavy capital is invested in the form of canals, drains, dams, water pumps, and barrages; the investment of skilled labour, commercial fertilizers, and pesticides is also great. Thus, despite multiple cropping, the yields of the land are exceptionally high. Strict crop rotation—in addition to government controls on the allocation of area to crops, on varieties planted, on the distribution of fertilizers and pesticides, and on marketing—contributes to the high productivity of agriculture.

Investments in agriculture

Unlike the situation in comparable developing countries, Egyptian agriculture has an overwhelmingly commercial rather than subsistence basis. Field crops contribute some three-fourths of the total value of Egypt's agricultural production, while the rest comes from livestock products, fruits and vegetables, and other specialty crops. Egypt has two seasons of cultivation, one for winter and another for summer crops. The main summer field crop is cotton, which occupies more than one-fifth of the season's arable land, absorbs much of the available labour, and represents a sizable portion of the value of exports. Egypt is the world's principal producer of long-staple cotton ($1\frac{1}{8}$ inches [2.85 centimetres] and longer), normally producing about one-third of the world crop, although total Egyptian production is only about 3 percent of all cotton produced in the world.

Among other principal field crops are corn (maize), rice, wheat, millet, and broad beans. Despite a considerable output, the cereal production in Egypt falls short of the country's total consumption; a substantial proportion of foreign exchange is spent annually on the import of cereals and milling products. Other important crops include sugarcane, alfalfa (lucerne), potatoes, and onions—the latter being normally an export item. Many varieties of fruit are grown, and some, such as citrus, are also exported.

Reclaimed land

In 1960–61 and 1968–69 about 896,100 acres were reclaimed. The total land reclaimed as a result of the Aswān High Dam project reached more than 1,000,000 acres by 1975, in addition to 700,000 acres converted from basin

(one crop a year) irrigation to perennial irrigation. During the same period, however, an area almost as large was lost to agriculture as industry and towns grew.

Egypt has been the scene of one of the most successful attempts at land reform. In 1952 a limit of 200 acres was imposed on individual ownership of land, and this was lowered to 100 acres in 1961 and to 50 acres in 1969. By 1975 less than one-eighth of the total cultivated area was held by owners with 50 acres or more. The success of Egyptian land reform is indicated by the substantial rise of land yields after 1952. This was partly the result of several complementary measures of agrarian reform, such as regulation of land tenure and rent control, that accompanied the redistribution of the land.

Following the construction of the Aswān High Dam, the Egyptian government encouraged the development of a thriving fishing industry. Construction of such projects as a fish farm and fishery complex at Lake Nasser have led to a considerable increase in the number of freshwater fish and in the size of the yearly total catch. At the same time, catches of sea fish in the waters off the Nile Delta have declined. This is thought to be a consequence of the change in the flow and character of Nile water after the construction of the Aswān High Dam.

Industry. The development of the manufacturing industry was handicapped by the policy of free trade imposed on Egypt from the middle of the 19th century until about 1930. Nationalism and World War II gave great impetus to the foundation of industrial projects that are largely agriculturally based and oriented toward import substitution. During the 1950s the country's manufacturing sector began to grow, and manufacturing and mining now account for a substantial portion of the gross domestic product.

Emphasis was placed on the development of heavy industry after a long-term agreement was signed with the Soviet Union in 1964. Another agreement with the Soviet Union, signed in 1970, provided aid for the expansion of the iron and steel complex at Ḥulwān; the establishment of a number of power-based industries, including an aluminum complex to utilize the power generated by the High Dam; and the electrification of the countryside. An ammonium nitrate fertilizer plant was opened in 1971, based on the gases generated in the coking unit of the steel mill at Ḥulwān. There is also a nitrate fertilizer plant at Aswān.

Egypt has made great achievements in increasing industrial production in such traditional industries as spinning and weaving, as well as in modern industries like engineering and iron and steel production. Food processing and the manufacture of chemical products also are important to the Egyptian economy.

Energy

Before the completion of the Aswān High Dam power station in 1970, the bulk of Egypt's electricity was generated in thermal stations using coal or diesel fuel, but some hydroelectric power was also generated by the old Aswān Dam. The 12 turbines of the High Dam power station have a capacity of about 2,000,000 kilowatts and are capable of producing 10,000,000,000 kilowatt-hours a year; the capacity of the thermal stations is about 45 percent of that of the High Dam. Transmission lines carry the current from Aswān to Cairo and to points farther north for use in urban centres and in manufacturing. The production of electric power from the High Dam has been limited, however, by the need to reconcile demands for power with the demands for irrigation water.

The bulk of Egypt's petroleum comes from the rich Morgan, Ramadan, and July fields (both onshore and offshore) in the Gulf of Suez, which are operated by the Gulf of Suez Petroleum Company, and from the Abū Rudays area of the Sinai on the Gulf of Suez. In cooperation with Phillips Petroleum Company, Egypt also extracts oil from fields at al-'Alamayn and Razzāq in the Western Desert. Active drilling for oil, involving several international interests, including those of the United States and several European nations, has continued in both the Eastern and the Western deserts.

Natural gas

In the process of searching for oil some significant natural gas deposits have been located. Phillips has located wells in the Abū Qīr area, northeast of Alexandria. A joint Egyptian-Italian gas discovery was made in the north Delta near Abū Māḍī in 1970; this was developed partly to supply a fertilizer plant and partly to fuel the industrial centres in the north and northwest Delta. In 1974 Abū Mādī became the first Egyptian gas field to begin production. Other natural gas fields are located in the Western Desert and the Gulf of Suez.

Egypt has several oil refineries, two of which are located at Suez. The first of Egypt's twin crude pipelines, linking the Gulf of Suez to the Mediterranean near Alexandria, was opened in 1977. This Suez–Mediterranean pipeline, known as Sumed, has an annual capacity for transmitting 80,000,000 tons of oil. The Sumed pipeline was financed by a consortium of Arab countries, primarily Saudi Arabia, Kuwait, and Egypt. In 1981 a crude oil pipeline was opened to link Ras Shuqir, on the Red Sea coast, with the refinery at Musturud, north of Cairo. An additional oil pipeline links Musturud with Alexandria.

Finance. The banking system of Egypt is centred on the Central Bank of Egypt, created in 1960 from the issue department of the National Bank of Egypt. In 1961 all banks operating in Egypt were nationalized, and their operations were concentrated in five commercial banks, in addition to the Central Bank, the government-sponsored Public Organization for Agricultural Credits and Co-operatives, the Development Industrial Bank, and three mortgage banks.

The government again reorganized the banking system in the early 1970s, merging some of the major banks and assigning special functions to each of the rest. Two new banks were created, and foreign banks were again permitted in the country as part of a program aimed at liberalizing the economy. Of particular interest were joint banking ventures between Egyptian and foreign banks. The stock exchanges at Cairo and Alexandria, which had been closed since the early 1960s, were reopened. The cotton exchanges in Cairo and Alexandria, which had also been closed, were replaced by a supervisory council responsible for regulating the cotton industry. In 1980 Egypt's first international bank was opened and a national investment bank was established.

The supply of money has, in general, followed the development of the economy; the authorities have aimed at tolerable increases in the price level, although since the 1973 war some prices have soared and inflation rates have risen sharply.

Egypt is a member of the International Monetary Fund. Since World War II the international liquidity of the Egyptian economy, including the Special Drawing Rights, added in 1970, has been depressed. In the late 1970s internal and external debts rose, mainly due to large government subsidies to the private sector. In the 1980s the government gradually introduced price increases on goods and services, with the goal of eventually reducing subsidies.

Imports and exports

Trade. Imports into Egypt average about one-third and exports about one-tenth of the gross domestic product. Since World War II exports have tended to fall short of imports. The trade deficit reached a peak in 1966 and was particularly sizable from 1960 to 1965 as expenditure on development rose. After the 1973 war there was a decided effort to restrict imports and stimulate exports, but this met with little success. The trade deficit continued to rise to record highs in the early and mid-1980s, largely because of the decline in revenue from petroleum exports and the increase in food imports.

Almost two-thirds of imports consist of raw materials, mineral and chemical products, and capital goods (machinery, electrical apparatus, and transport equipment), more than one-fourth are foodstuffs, and the remainder are other consumer goods. More than one-half of the exports by value consist of petroleum and petroleum products, followed by raw cotton, cotton yarn, and fabrics. Raw materials, mineral and chemical products, and capital goods are also exported. Among agricultural exports are rice, onions, garlic, and citrus fruit.

Italy, France, and the Soviet Union are among Egypt's largest markets. The United States, however, is the major source of Egypt's imports, followed by West Germany, Italy, and France.

The economic boycott by other Arab states, which resulted from the 1979 peace treaty between Israel and Egypt, did not have a serious effect on Egypt's economy. In the early and mid-1980s Egypt's revenue fluctuated, however, in response to changes in oil sales and tourist revenue, and the country continued to have deficits in its foreign-trade balance. The deficit has been financed by international borrowing (primarily from the International Monetary Fund), transfers from Arab oil-producing countries, revenue from expatriate remittances, Suez Canal revenue, and changes in foreign assets and liabilities.

(E.I.U./D.H./C.G.S.)

Transportation. Almost the entire communication system is state owned. It is adequate in terms of coverage, but stresses arise from excessive usage. The main patterns of transport flow reflect the topographical configuration of the country—that is to say, they follow the north–south course of the Nile, run along the narrow coastal plain of the Mediterranean Sea, and expand into a more complex system in the Delta.

Road network. About half of Egypt's total road network is paved. Rural roads are of dried mud, usually following the lines of the irrigation canals; many of the desert roads are little more than tracks. The Cairo–Alexandria highway runs via Banhā, Ṭanṭā, and Damanhūr. The alternate desert road to Cairo from Alexandria has been extensively improved, and a good road links Alexandria with Libya by way of Maṭrūḥ on the Mediterranean coast. There are paved roads between Cairo and al-Fayyūm, and good roads connect the various Delta and Suez Canal towns. A paved road parallels the Nile from Cairo south to Aswān, and another paved road runs from Asyūṭ to al-Khārijah and ad-Dākhilah in the Western Desert. The coastal Red Sea route to Marsā al-'Alām is poorly paved, as are the connecting sections inland.

Railways. Railways connect Cairo with Alexandria and with the Delta and canal towns and also run southward to Aswān and the High Dam. Branch lines connect Cairo with al-Fayyūm and Alexandria with Maṭrūḥ. A network of light railways connects the Fayyūm area and the Delta villages with the main lines. Diesel-driven trains operate along the main lines; electric lines connect Cairo with the suburbs of Ḥulwān and Heliopolis.

Navigable waterways. The Suez Canal, closed in 1967, was reopened in 1975; it serves as a major link between the Mediterranean and Red seas. The Nile and its associated navigable canals provide an important means of transportation, primarily for heavy goods. There are approximately 2,000 miles of navigable waterways—about one-half of this total on the Nile itself, which is navigable throughout its length. The inland-waterway freight fleet consists of tugs, motorized barges, towed barges, and flat-bottomed feluccas (two- or three-masted lateen-rigged sailing ships).

Rivers and canals

Ports and shipping. In spite of its long coastline, Egypt has only three ports of any significance—Alexandria, Port Said, and Suez. Alexandria, with a fine natural harbour, handles most of the country's imports and exports, as well as the bulk of passenger traffic. Port Said, at the northern entrance to the Suez Canal, lacks the berthing and loading facilities of Alexandria. Suez's main function is that of an entry port for petroleum and minerals from the Egyptian Red Sea coast and for goods from the Far East.

Air transport. Cairo is an important communication centre for world air routes. The enlarged airport at Heliopolis, with its modern terminal building, is used by major international airlines, as is Nuzhah airport at Alexandria.

The national airline, Egypt Air, runs external services throughout the Middle East, as well as to Europe, North America, Africa, and the Far East; it also operates a domestic air service.

GOVERNMENT AND SOCIAL CONDITIONS

Government. Before the 1952 revolution, Egypt was a constitutional monarchy; the 1923 constitution, which followed the declaration of the end of the British protectorate, stated that Egypt was an independent sovereign Islāmic state with Arabic as its language and provided for a representative parliament. This constitution was abolished in 1952, political parties were dissolved in 1953, and a new constitution was introduced in 1956. The Republic of Egypt was declared. Between 1958 and 1961 Egypt and Syria were merged into one state, called the United Arab Republic; the name was retained by Egypt upon Syria's secession in 1961. The National Union, organized in 1957 in place of the political parties abolished in 1953, became the Arab Socialist Union (ASU) in 1962.

In 1971 Egypt, Libya, and Syria agreed to the establishment of the Confederation of Arab Republics. A draft constitution was agreed to by the heads of state of each country and was approved by referenda in each of the three member states. The capital of the confederation was Cairo. In 1979, however, deteriorating relations between Egypt and other Arab nations led to the end of the confederation; following the signing of the Egyptian-Israeli peace treaty, most Arab economic ties with Egypt also were suspended.

On Sept. 11, 1971, a new constitution for Egypt was approved by referendum. It proclaimed the Arab Republic of Egypt to be "a democratic, socialist state" with Islām as its state religion and Arabic as its national language. It recognized three types of ownership—public, cooperative, and private. It guaranteed the equality of all Egyptians before the law and their protection against arbitrary intervention in the processes of law. It also affirmed the rights to peaceful assembly, education, and health and social security, and the right to organize into associations or unions and to vote.

According to the constitution and its subsequent amendments, the president of the republic is the head of state, and, together with the Cabinet, constitutes the executive authority. The president must be Egyptian, born of Egyptian parents, and not less than 40 years old. The presidential term is six years and may be extended to additional terms. The president has the power to appoint and dismiss one or more vice presidents, the prime minister, ministers, and deputy ministers. The legislative body is composed of the People's Assembly, which nominates the presidential candidate by a two-thirds majority. The candidate is then confirmed by national plebiscite.

The president is the supreme commander of the armed forces and has the right to grant amnesty and reduce sentence, the power to appoint civil and military officials and to dismiss them in a manner prescribed by the law, and the authority to call a referendum of the people on matters of supreme importance. The president can, in exceptional cases and by investiture of the assembly, issue decrees having the force of law—but only for a defined period of time.

Legislative power resides in the People's Assembly, which is composed of 448 elected members, some of whom must be women, and 10 additional members appointed by the president. The assembly is elected, under a complex system of proportional representation, for a five-year term. All males 18 years of age and older are required to vote, as well as all women on the register of voters. The president convenes and closes the sessions of the People's Assembly.

The People's Assembly

The People's Assembly's main function is to approve policy. Its members must ratify all laws and examine and approve the national budget. It also approves the program of each newly appointed Cabinet. Should it withdraw its confidence from the Cabinet or any of its members, that person is required to resign. The president cannot dissolve the assembly except under special circumstances and after a vote of approval by a people's referendum. Elections for a new assembly must be held within 60 days of dissolution.

The constitution also provides for a judiciary, independent of other authorities, whose functions and authority are governed by special legislation. A Council of National Defence, presided over by the president of the republic, is responsible for matters relating to security and defense.

Local government and administration. Until 1960, government administration was highly centralized; in that year, however, the local-government administrative system was established to promote decentralization and greater citizen participation in local government.

The 1960 Local Administration Law provides for three levels of local administration—the *muḥāfaẓāt* (gover-

Governorates, districts, and villages

norates), the *markaz* (districts or counties), and the *qarīyah* (villages). The structure combines features of both local administration and local self-government. There are two councils at each administrative level: a mostly elected people's council and an appointed executive council. Although these councils exercise broad legislative powers, they are controlled by the central government.

The country is divided into 26 *muḥāfaẓāt*. Five cities—Cairo, Alexandria, Ismailia, Port Said, and Suez—have *muḥāfaẓah* status. The governor is appointed and can be dismissed by the president of the republic. He is the highest executive authority in the *muḥāfaẓah,* has administrative authority over all government personnel except judges in his *muḥāfaẓah,* and is responsible for implementing policy.

The *muḥāfaẓah* council is composed of a majority of elected members. Although it has not been possible in practice, according to law at least one-half of the members of the *muḥāfaẓah* council are to be farmers and workers. The town or district councils and the village councils are established on the same principles as those underlying the *muḥāfaẓah* councils.

The local councils perform a wide variety of functions in education, health, public utilities, housing, agriculture, and communications; they are also responsible for promoting the cooperative movement and for implementing parts of the national plan. Local councils obtain their funds from national revenue, a tax on buildings and lands within the *muḥāfaẓah,* miscellaneous local taxes or fees, profits from public utilities and commercial enterprises, and national subsidies, grants, and loans.

The political process. After 1962 all popular participation and representation in the political process was through the Arab Socialist Union. In 1976, however, the ASU lost its status as the sole legal political organization, and other political parties soon formed; their right to exist was recognized by a law adopted in June 1977. The ASU was abolished by constitutional amendment in 1980.

The Arab Socialist Union

The National Democratic Party (NDP), formed by Pres. Anwar el-Sādāt in 1978, serves as the official government party and holds a majority of seats in the People's Assembly. The left-wing opposition is the National Progressive Unionist Party and the Socialist Labour Party. The prerevolutionary Wafd Party has been re-formed, and one religious party, the Umma, has been licensed. Officially unrepresented are the Communists, extreme religious groups, and avowed Nasserists.

Justice. The Egyptian constitution emphasizes the independent nature of the judiciary. There is to be no external interference with the due processes of justice. Judges are subject to no authority other than the law; they cannot be dismissed and are disciplined in the manner prescribed by law. Judges are appointed by the state, with the prior approval of the Supreme Judicial Council under the chairmanship of the president. The council is also responsible for the affairs of all judicial bodies; its composition and special functions are specified by law.

The courts

The court structure can be regarded as falling into four categories, each of which has a civil and criminal division. These courts of general jurisdiction include district tribunals, tribunals of the first instance, courts of appeal, and the Court of Cassation. Court sessions are public, except where consideration of matters of public order or decency decides otherwise. Sentence is passed in open session.

In addition, there are special courts, such as military courts and courts of public security—the latter dealing with crimes against the well-being or security of the state. The Council of State is a separate judicial body, dealing especially with administrative disputes and disciplinary actions. The Supreme Constitutional Court in Cairo is the highest court in Egypt. Its functions include judicial review of the constitutionality of laws and regulations and the resolution of judicial conflicts among the courts.

Law enforcement. The Ministry of the Interior has direct control and supervision over all police and security functions at the *muḥāfaẓah,* district, and village levels. At the central level, the deputy minister for public security is responsible for general security, emigration, passports, port security, criminal investigation, ministerial guards, and emergency services. The deputy minister for special police is responsible for civil defense, traffic, prison administration, tourist police, and police transport and communications.

Education. At the end of the 19th century there were only three secondary and nine "higher" schools in Egypt; the educational structure continued to be based on the *kuttābs*, or Qur'ān schools. In 1916 the latter were turned into elementary schools, and in 1923 a law was passed providing free compulsory education between the ages of seven and 12. A sharp increase in the annual budgetary allocation devoted to education occurred after World War II. Following the revolution of 1952, educational progress already achieved was accelerated and was accompanied by both the Egyptianization and Arabicization of the educational system. One of the most significant features of this progress has been the spread of women's education. By the late 1970s almost one-third of the students attending university were women. Women are no longer confined to the home; many fields of employment, including the professions and even politics, are now open to them. A further result of the expansion of education has been the emergence of an intellectual elite and the growth of a middle class, consisting of members of the professions, government officials, and businessmen. In spite of the rapid advance in the provision of education services, however, illiteracy has remained relatively high.

There are three stages of state general education—primary (six years), preparatory (three years), and secondary (three years). Primary education between the ages of six and 12 is compulsory. Pupils who are successful in examinations have the opportunity to continue their education first at the preparatory and then at the secondary level. There are two types of secondary school, general and technical; most technical schools are either commercial, agricultural, or industrial.

Alongside the Ministry of Education's system of general education, there is that provided by the institutes associated with al-Azhar University, centred on al-Azhar Mosque in the medieval quarter of Cairo. Al-Azhar has been a teaching centre for the entire Muslim world for nearly a millennium. Instruction is given at levels equivalent to those of the state schools, but in order to allow for greater emphasis on traditional Islāmic subjects, the duration of training is lengthened by one year at the preparatory stage and two at the secondary. A large-scale modernization of the college-level curriculum, making it comparable to those of other state universities, has been carried out since 1961.

In the 1950s there were almost 300 foreign schools in Egypt, the majority of them French; many of these have since become, to varying degrees, Egyptianized. Pupils who attend these schools, at all levels, sit for the same state certificate examinations as those in the normal state system.

The universities

The major state universities are Cairo, Alexandria, 'Ayn Shams, and Asyūṭ. In addition to the state university system, there is one private university, the American University in Cairo.

There are many institutes of higher learning, excluding institutes attached to universities or affiliated to the Ministry of Culture—such as the Institute of Dramatic Arts, the Cinema Institute, and the Institute of Ballet. These institutes specialize in commerce, industry, agriculture, the arts, physical culture, social service, domestic economy, and languages. Courses of study lead to a degree.

Health and welfare. The budget of the Ministry of Health has reflected a steadily increasing expenditure on public-health programs, and the numbers of government health centres, beds in public hospitals, doctors, and dentists have increased dramatically.

An important aspect of this development has been the expansion of health facilities in the rural areas of the country. In 1953 the government introduced what are termed combined service units; these differ from health centres in that they combine the functions of health centre, school, social-welfare unit, and agricultural extension services. In addition, rural health units further extend the health services available in rural communities. Each unit

is operated by a team of seven or eight people, including one physician.

Well-trained physicians and specialists are available in large numbers in the cities and larger towns. The medical profession has prestige, and only the better qualified high school graduates are accepted into medical schools.

Public health campaigns

Significant efforts have been made to promote preventive medicine. Compulsory vaccination against smallpox, diphtheria, tuberculosis, and poliomyelitis is enforced for all infants during their first two years. Schistosomiasis, a parasitic disease that is widespread among the rural population, presents a serious health problem. All health centres offer treatment against it, but reinfection can easily occur. Epidemics of malaria have been eliminated, but the disease still exists in endemic form, mainly in southern Egypt. Treatment for malaria is provided at all health centres, and the spraying of houses in mosquito-breeding areas is carried out regularly. Attention has also been given to the problem of tuberculosis; centres have been established in every *muḥāfaẓah,* and mass X-ray and immunization campaigns have been carried out.

The government has attempted to socialize medicine through such measures as the nationalization and control of pharmaceutical industries, the nationalization of hospitals run by private organizations and associations, and expanded health insurance. A health insurance law was passed in 1964; it provides for compulsory health insurance for workers in firms employing more than 100 persons, as well as for all governmental and public employees.

Housing. Egypt has faced a serious urban housing shortage since World War II. The situation subsequently became aggravated by increased immigration from rural to urban areas, resulting in extreme urban overcrowding.

Housing construction

Although there is considerable concern over the housing problem, the combined efforts of both public and private sectors have been unable to meet the growing demand. Between 1970 and 1980, for example, approximately 300,000 housing units were built; this represented an increase of more than one-fourth of the total number of housing units. The increase in the urban population, however, was estimated at more than 40 percent during the same period; *i.e.,* for every new housing unit built, 13 persons were added to the urban population.

In the rural areas villagers build their own houses at little cost with the materials available. The government has experimented in aiding self-help projects with state loans. Ambitious rural housing projects have been carried out on newly reclaimed land: entire villages with all the necessary utilities have been built.

CULTURAL LIFE

In spite of the many ancient civilizations with which it has come into contact, Egypt unquestionably belongs to a sociocultural tradition that is Arabic and Islāmic. This tradition remains a constant factor in determining Egyptian views both of itself and of the world.

The story of the cultural development of modern Egypt is, in essence, that of the response of this traditional system to the intrusion into it, at first by conquest and later by the penetration of ideas, of the alien and materially superior civilization of the West. The response covered a broad spectrum—from the rejection of new ideas and reversion to traditionalism through self-examination and reform to an uncritical acceptance of new concepts and the values that went with them. The result has been the emergence of a cultural identity devoid of self-consciousness, which has assimilated much that is new, while remaining distinctively Egyptian. The process is to be seen at work in all branches of contemporary culture.

The state of the arts. The impact of the West is one of the recurring themes in the modern Egyptian novel, as in Tawfīq al-Ḥakīm's *'Usfūr min ash-Sharq* ("The Bird from the East") and Yaḥyā Ḥaqqī's novella *Qindīl Umm Hāshim* ("The Lamp of Umm Hāshim"). A further theme is that of the Egyptian countryside—romantically handled at first, as in Muḥammad Ḥusayn Haykal's *Zaynab,* and later realistically, as in 'Abd al-Raḥmān ash-Sharqāwī's *al-Arḍ* (*The Land*) and *al-Fallāḥ* ("The Peasant") and in Yūsuf Idrīs' *al-Ḥarām* ("The Forbidden"). A Dickensian capacity to catch the colour of life among the urban poor is a characteristic quality of the early and middle work of Egypt's greatest modern novelist, Najīb Maḥfūẓ, notably in *Zuqāq al-Midaqq* (*Midaq Alley*).

The modern theatre in Egypt is a European importation—the first Arabic-speaking plays were performed in 1870. Two dramatists, both born at the turn of the century, have dominated its development—Maḥmūd Taymūr and Tawfīq al-Ḥakīm. The latter, a versatile and cerebral playwright, has reflected in his themes not only the development of the modern theatre but also, in embryo, the cultural and social history of modern Egypt. The changes in Egyptian society are reflected in the themes adopted by younger dramatists.

Film-making in Egypt

There is a relatively long tradition of filmmaking in Egypt going back to World War I, but it was the founding of Miṣr Studios in 1934 that stimulated the growth of the Arabic-speaking cinema. Modern Egyptian films are shown to audiences throughout the Arab world and are also distributed in Asian and African countries. The industry is both privately and state owned—there are many private film-production companies, as well as the Ministry of Culture's Egyptian General Cinema Corporation.

Contemporary Egyptian music embraces indigenous folk music, traditional Arabic music, and Western-style music. The revival of traditional Arabic music, both vocal and instrumental, owes much to state sponsorship. Popular Arabic music consists of a blend of classical Arabic music, folk songs, and Western music. Muḥammad 'Abd al-Wahhāb has been one of the leading figures in the development of this genre, as both composer and singer. Umm Kulthūm was the leading vocalist not only of Egypt but also of the whole Arab world for almost 50 years. Western-style music has been a familiar component in Egyptian musical culture since the 19th century. Pioneers such as Yūsuf Greiss and Abū Bakr Khayrat succeeded in incorporating Arabic elements to give a national colouring to their Western-style compositions.

A return to folklore as a source of inspiration for the arts is a generalized phenomenon in modern Egyptian culture. It has resulted in a revived interest in traditional crafts, in the collection of folk music, and the maintaining, with government sponsorship, of two folk-dance ensembles—the Riḍā Troupe and the National Folk Dance Ensemble. In the plastic arts the highly original use of local themes is particularly striking. An active school of Egyptian painting and sculpture has emerged.

Cultural institutions. The oldest learned academy in Egypt, the Institut d'Égypte, was founded in 1859, but its antecedents go back to the institute established by Napoleon in 1798. The Academy of the Arabic Language, founded in 1932 and presided over by the veteran educator Ṭaha Ḥusayn, became, in terms of prestige and influence, one of the most important cultural institutions in Egypt. Linked to the Ministry of Culture, it enjoys a large measure of autonomy, guaranteed by its own charter. Also attached to the Ministry of Culture is the Higher Council for Arts, Letters, and the Social Sciences. Intended as a consultative body on cultural matters, the Higher Council is also a means of channeling state patronage.

The learned societies

Learned societies in Egypt support a wide variety of interests—including the physical and natural sciences, medicine, agriculture, the humanities, and the social sciences. Increased government concern with research, especially in science and technology, was reflected in the founding of the National Research Centre, where laboratory work in both pure and applied science began in 1956, and of the Atomic Energy Establishment, in 1957. In addition, there are many specialized research institutes in the country.

Most of the learned societies and research institutes have library collections of their own. In addition to large collections at the universities, the municipalities of Alexandria, al-Manṣūrah, and Ṭanṭā maintain libraries. There is also a central public library in each *muḥāfaẓah,* with branches in small towns and service points in the villages. The Ministry of Culture is responsible for the Egyptian National Library (Dār al-Kutub) and the National Archives, both in Cairo, and the Public Libraries Administration. The

Egyptian National Library, which has a large collection of printed materials, is also a centre for the collection and preservation of manuscripts.

The Ministry of Culture is also responsible, through its department of antiquities, for the Egyptian Museum, the Coptic Museum, and the Museum of Islāmic Art, all in Cairo; the Greco-Roman Museum in Alexandria; and for other institutions, including fine-arts museums such as the Mukhtār Museum, the Nājī Museum, and the Museum of Modern Art, all in Cairo, and the Museum of Fine Arts in Alexandria.

All newspapers and magazines in Egypt are subject to supervision through the government's Supreme Press Council. Daily newspapers include the long-established *al-Ahram,* published in Cairo, and other Arabic-language papers, together with daily English-language and French-language newspapers. The government owns and operates the Egyptian Radio and Television Corporation, which provides programs in a variety of languages. Cairo is considered to be the largest centre of publishing in the Middle East.

For statistical data on the land and people of Egypt, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (L.S.El H./Ma.J./D.H./C.G.S.)

History

INTRODUCTION TO ANCIENT EGYPTIAN CIVILIZATION

Life in ancient Egypt. Ancient Egypt can be thought of as an oasis in the desert of northeast Africa, dependent on the annual inundation of the Nile to support its agricultural population. The country's chief wealth came from the fertile floodplain of the Nile Valley, where the river flows between bands of limestone hills, and the Nile Delta, in which it fans into several branches north of modern Cairo. Between the floodplain and the hills is a variable band of low desert, which supported a certain amount of game. The Nile was Egypt's sole transportation artery.

The Nile floodplain

The First Cataract at Aswān, where the riverbed is turned into rapids by a belt of granite, was the country's only well-defined boundary within a populated area. To the south lay the far less hospitable area of Nubia, in which the river flowed through low sandstone hills that left a very narrow strip of cultivable land. Nubia was significant for Egypt's periodic southward expansion and for access to products from farther south. West of the Nile was the arid Sahara, broken by a chain of oases some 125–185 miles (about 200–300 kilometres) from the river and lacking in all other resources except for a few minerals. The eastern desert, between the Nile and the Red Sea, was more important, for it supported a small nomadic population and desert game, contained numerous mineral deposits including gold, and was the route to the Red Sea.

To the northeast was the Isthmus of Suez. It offered the principal route for contact with Sinai, from which came turquoise and possibly copper, and with western Asia, Egypt's most important area of cultural interaction, from which were received stimuli for technical development and cultivars for crops. Immigrants and ultimately invaders crossed the Isthmus into Egypt, attracted by the country's stability and prosperity. From the late 2nd millennium BC on, numerous attacks were made by land and sea along the eastern Mediterranean coast.

At first, relatively little cultural contact came by way of the Mediterranean Sea, but from an early date Egypt maintained trading relations with the Lebanese port of Byblos (modern Jubayl). Egypt needed few imports to maintain basic standards of living, but good timber was essential and not available within the country, so it usually was obtained from Lebanon. Minerals such as obsidian and lapis lazuli were imported from as far afield as Anatolia and Afghanistan.

Agriculture

Agriculture centred on the cultivation of cereal crops, chiefly emmer wheat (*triticum dicoccum*) and barley (*hordeum vulgare*). The fertility of the land and general predictability of the inundation ensured very high productivity from a single annual crop. This productivity made it possible to store large surpluses against crop failures and also formed the chief basis of Egyptian wealth, which was, until the creation of the large empires of the 1st millennium BC, the greatest of any state in the ancient Near East.

Irrigation was achieved by simple means and multiple cropping was not feasible until much later times, except perhaps in the lakeside area of Fayyūm. As the river deposited alluvial silt, raising the level of the floodplain, and land was reclaimed from marsh, the area available for cultivation in the Nile Valley and Delta increased, while pastoralism declined slowly. In addition to grain crops, fruit and vegetables were important, the latter being irrigated year-round in small plots; and fish was vital to the diet. Papyrus, which grew abundantly in marshes, was gathered wild and in later times was cultivated. It may have been used as a food crop; and it certainly was used to make rope, matting, and sandals. Above all it provided the characteristic Egyptian writing material, which, with cereals, was the country's chief export in Late Period Egyptian and then Greco-Roman times.

After the introduction of cultivated cereal crops, meat was eaten mainly by the wealthy. Domesticated animals lost much of their significance for nutrition, but they retained great cultural importance and practical value. Cattle may have been domesticated in northeastern Africa. The Egyptians kept many as draft animals and for their various products, showing some of the interest in breeds and individuals that is found to this day in the Sudan and eastern Africa. The donkey, which was the principal transport animal (the camel did not become common until Roman times), was probably domesticated in the region. The native Egyptian breed of sheep became extinct in the 2nd millennium BC and was replaced by an Asiatic breed. Wool was rarely used, so that sheep were primarily a source of meat. Goats were more numerous than sheep and were commonly depicted browsing on tree foliage. Pigs, although subject to some sort of taboo, were raised and eaten. Ducks and geese were kept for food, and many of the vast numbers of wild and migratory birds found in Egypt were hunted and trapped. Desert game, principally various species of antelope and ibex, were hunted by the elite; it was a royal privilege to hunt lions and wild cattle. Pets included dogs, which were also used for hunting; cats (domesticated in Egypt); and monkeys. In addition, the Egyptians had a great interest in, and knowledge of, most species of mammals, birds, reptiles, and fish in their environment.

Most Egyptians were probably descended from settlers who came to the Nile Valley in prehistoric times, with increase coming through natural fertility. In various periods there were immigrants from Nubia, Libya, and especially the Near East. They were historically significant and may have contributed to population increase, but their numbers are unknown. Most people lived in villages and towns in the Nile Valley and Delta. Dwellings were normally built of mud brick and have long since disappeared beneath the rising water table, thereby obliterating evidence for settlement patterns. In antiquity, as now, the most favoured location of settlements was on slightly raised ground near the riverbank, where transport and water were easily available and flooding was unlikely. Until the 1st millennium BC Egypt was not urbanized to the same extent as Mesopotamia. Instead, a few centres, notably Memphis and Thebes, attracted population and particularly the elite, while the rest of the people were relatively evenly spread over the land. The size of the population has been estimated as rising from between 1,000,000 and 1,500,000 in the 3rd millennium BC to perhaps twice as many in the late 2nd millennium and 1st millennium BC. (Much higher levels of population were reached in Greco-Roman times.)

Ownership of land

Nearly all of the people were engaged in agriculture and were probably tied to the land. All the land belonged in theory to the king, although in practice those living on it could not easily be removed and some categories of land could be bought and sold. Land was assigned to high officials to provide them with an income, and most categories of land paid substantial dues to the state, which had a strong interest in keeping it in agricultural use. Abandoned land was taken back into state ownership and reassigned for cultivation. The people who lived on and worked the

land were not free to leave and were obliged to work it, but they were not slaves; most paid a proportion of their produce to major officials. Free citizens who worked the land on their own behalf did emerge; terms used for them tended originally to refer to poor people, although they were probably not in fact poor. Slavery was never very common, being restricted to captives and foreigners or to people who were forced by poverty or debt to sell themselves into service. Slaves sometimes were even married by members of their owners' families, so that in the long term those belonging to households tended to be assimilated into free society. In the New Kingdom (from about 1539 to 1075 BC), large numbers of captive slaves were acquired by major state institutions or incorporated into the army. Punitive treatment of foreign slaves or of native fugitives from their obligations included forced labour, exile (in, for example, the oases of the western desert), or compulsory enlistment in dangerous mining expeditions. Even nonpunitive employment such as quarrying in the desert was hazardous. The official record of one expedition shows a mortality rate of more than 10 percent.

Just as the Egyptians optimized agricultural production with simple means, their crafts and techniques, many of which originally came from Asia, were raised to extraordinary levels of perfection. The Egyptians' most striking technical achievement, massive stone building, also exploited the potential of a centralized state to mobilize a huge labour force, which was made available by efficient agricultural practices. Some of the technical and organizational skills involved were remarkable. The construction of the great pyramids of the 4th dynasty (*c.* 2575–*c.* 2465 BC) has yet to be fully explained and would be a major challenge to this day. This expenditure of skill contrasts with sparse evidence for an essentially neolithic way of living for the rural population of the time, while the use of flint tools persisted even in urban environments at least until the late 2nd millennium BC. Metal was correspondingly scarce, much of it being used for prestige rather than everyday purposes.

Family, marriage, and the role of women

In urban and elite contexts the Egyptian ideal was the nuclear family, but on the land and outside the central ruling group there is evidence for extended families. Egyptians were monogamous, and the choice of partners in marriage, for which no formal ceremony or legal sanction is known, did not follow a set pattern. Consanguineous marriage was not practiced during the Dynastic Period, except for the occasional marriage of a brother and sister within the royal family, and the practice may have been open only to kings or heirs to the throne. Divorce was in theory easy, but it was very costly. Women had a legal status only marginally inferior to that of men. They could own and dispose of property in their own right, and they could initiate divorce and other legal proceedings. They hardly ever held administrative office but increasingly were involved in religious cults as priestesses or "chantresses." Elite married women held the title "Mistress of the House," the precise significance of which is unknown. Lower down the social scale they probably worked on the land as well as in the house.

The uneven distribution of wealth, labour, and technology was related to the only partly urban character of society, especially in the 3rd millennium BC. The country's resources were not fed into numerous provincial towns but instead were concentrated to great effect around the capital—itself a dispersed string of settlements rather than a city—and focused on the central figure in society, the king. In the 3rd and early 2nd millennia the elite ideal, expressed in the decoration of private tombs, was manorial and rural. Not until much later did Egyptians have pronouncedly urban values.

The king and ideology: administration, art, and writing. In official terms, Egyptian society consisted of a descending hierarchy of the gods, the king, the dead, and humanity (by which was understood chiefly the Egyptians). Of these groups, only the king was single, and hence he was individually more prominent than any of the others. A text that summarizes the king's role states that he "is on earth for ever and ever, judging mankind and propitiating the gods, and setting order [*ma'at,* a central concept] in place of disorder. He gives offerings to the gods and mortuary offerings to the spirits [the blessed dead]." The king was a god, but not in any simple or unqualified sense. His divinity accrued to him from his office and was reaffirmed through rituals, but it was vastly inferior to that of major gods; he was god rather than man by virtue of his potential, which was immeasurably greater than that of any human being. To humanity, he manifested the gods on earth, a conception that was elaborated in a complex web of metaphor and doctrine; less directly, he represented humanity to the gods. The text quoted above also gives great prominence to the dead, for whom the living performed a cult and who could intervene in human affairs; in many periods the chief visible expenditure and focus of display of nonroyal individuals, as of the king, was on provision for the tomb and the next world. Egyptian kings are commonly called pharaohs, following the usage of the Old Testament. The term pharaoh, however, is derived from the Egyptian *per 'aa* ("great estate") and goes back to the designation of the royal palace as an institution. This term for palace was used increasingly from about 1400 BC as a way of referring to the living king; in earlier times it was rare.

The king's relation to the gods

Rules of succession to the kingship are poorly understood. The common conception that the heir to the throne had to marry his predecessor's oldest daughter has been disproved; kingship did not pass through the female line. The choice of queen seems to have been free: often the queen was a close relative of the king, but she also might be unrelated to him. In the New Kingdom, for which evidence is abundant, each king had a queen with distinctive titles, as well as a number of minor wives.

Sons of the queen seem to have been the preferred successors to the throne, but other sons could also become king. In many cases the successor was the eldest (surviving) son, and such a pattern of inheritance agrees with more general Egyptian values, but often he was some other relative, or was completely unrelated. New Kingdom texts depict, after the event, how kings were appointed heirs either by their predecessors or by divine oracles, and such may have been the pattern when there was no clear successor. From the middle of the 5th dynasty (*c.* 2450 BC) to the 19th (1292–1190 BC) there is no certain attestation of a prince in the reign of his brother; rival claimants, therefore, must have been eliminated or silenced after one of them had succeeded. Dissent and conflict are suppressed from public sources. From the Late Period (664–332 BC), when sources are more diverse and patterns less rigid, numerous usurpations and interruptions to the succession are known; they probably had many forerunners.

The king's position changed gradually from that of an absolute monarch at the centre of a small ruling group who were mostly his kin to that of the head of a bureaucratic state—in which his rule was still absolute—based on officeholding and, in theory, on free competition and merit. By the 5th dynasty, fixed institutions were added to the force of tradition and the regulation of personal contact as brakes on autocracy, but the charismatic and superhuman power of the king remained vital.

The bureaucratic elite

The elite of administrative officeholders received their positions and commissions from the king, whose general role as judge over humanity they put into effect. They commemorated their own justice and concern for others, especially their inferiors, and recorded their own exploits and ideal conduct of life in inscriptions for others to see. Thus the position of the elite was affirmed by reference to the king, to their prestige among their peers, and to their conduct toward their subordinates, justifying to some extent the fact that they—and still more the king—appropriated much of the country's surplus production for their own benefit.

These attitudes and their potential dissemination through society counterbalanced inequality, but how far they were accepted cannot be known. The core group of wealthy officeholders numbered at most a few hundred, and the administrative class of minor officials and scribes, most of whom could not afford to leave memorials or inscriptions, perhaps 5,000. With their dependents, these two groups formed perhaps 5 percent of the early population. Monu-

ments and inscriptions commemorated no more than one in a thousand people.

According to royal ideology, the king appointed the elite on the basis of merit, and in ancient conditions of high mortality the elite had to be open to recruits from outside. In addition, royal caprice resulted in many falls from favour, especially in the 18th dynasty (1539–1292 BC). There was, however, also an ideal that a son should succeed his father. In periods of weak central control this principle predominated, and in the Late Period the whole society became more rigid and stratified.

Writing was a major instrument in the centralization of the Egyptian state and its self-presentation. The two basic forms of writing, hieroglyphs, which were used for monuments and display, and the cursive form known as hieratic, were invented at much the same time in late predynastic Egypt (*c.* 3000 BC). Writing was chiefly used for administration and until about 2650 BC no continuous texts were recorded; the only literary texts written down before the early Middle Kingdom (*c.* 1950 BC) seem to have been lists of important traditional information and possibly medical treatises. The use and potential of writing were restricted both by the rate of literacy, which was probably well below 1 percent, and expectations of what writing might do. Hieroglyphic writing was publicly identified with Egypt. Perhaps because of this association with a single powerful state, its language, and its culture, Egyptian writing was seldom adapted to write other languages; in this it contrasts with the cuneiform script of the relatively uncentralized, multilingual Mesopotamia. Nonetheless, Egyptian hieroglyphs probably served in the middle of the 2nd millennium BC as the model from which the alphabet, ultimately the most widespread of all writing systems, evolved.

Use of hieroglyphs

The dominant visible legacy of ancient Egypt is in works of architecture and representational art. Until the Middle Kingdom, most of these were mortuary: royal tomb complexes, including pyramids and mortuary temples, and private tombs. There were also temples dedicated to the cult of the gods throughout the country, but most of these were modest structures. From the beginning of the New Kingdom (*c.* 1539 BC), temples of the gods became the principal monuments; royal palaces and private houses, which are very little known, were less important. Temples and tombs were ideally executed in stone with relief decoration on their walls and were filled with stone and wooden statuary, inscribed and decorated stelae (freestanding small stone monuments), and, in their inner areas, composite works of art in precious materials. The design of the monuments and their decoration goes back in essence to the beginning of the historical period and presents an ideal, sanctified cosmos. Little in it is related to the everyday world and, except in palaces, works of art may have been rare outside temples and tombs. Decoration may record real historical events, rituals, or the official titles and careers of individuals, but its prime aim is the more general assertion of values, and the information presented must be evaluated for its plausibility and compared with other evidence. Some of the events depicted in relief on royal monuments were certainly fictitious.

Temples and tombs and their decoration

The highly distinctive Egyptian method of rendering nature and artistic style were also creations of early times and can be seen in most works of Egyptian art. In content, these are hierarchically ordered so that the most important figures, the gods and the king, are shown together, while before the New Kingdom gods seldom occur in the same context as humanity. The decoration of a nonroyal tomb characteristically shows the tomb's owner with his subordinates, who administer his land and present him with its produce. The tomb owner is also typically depicted hunting in the marshes, a favourite pastime of the elite that may additionally symbolize passage into the next world. The king and the gods are absent in nonroyal tombs, and overtly religious matter is restricted to rare scenes of mortuary rituals and journeys and to textual formulas. Temple reliefs, in which king and gods occur freely, show the king defeating his enemies, hunting, and especially offering to the gods, who in turn confer benefits upon him. Human beings are present at most as minor figures supporting the king. On both royal and nonroyal monuments an ideal world is represented in which all are beautiful and everything goes well; only minor figures may have physical imperfections.

This artistic presentation of values originated at the same time as writing, but before the latter could record continuous texts or complex statements. Some of the earliest continuous texts of the 4th and 5th dynasties show an awareness of an ideal past that the present could only aspire to emulate. A few "biographies" of officials allude to strife, but more nuanced discussion occurs first in literary texts of the Middle Kingdom. The texts consist of stories, dialogues, lamentations, and especially instructions on how to live a good life, and they supply a rich commentary on the more one-dimensional rhetoric of public inscriptions. Literary works were written in all the main later phases of the Egyptian language—Middle Egyptian; the "classical" form of the Middle and New kingdoms, continuing in copies and inscriptions into Roman times; Late Egyptian, from the 19th dynasty to about 700 BC; and demotic (texts from the 4th century BC to the 3rd century AD)—but many of the finest and most complex are among the earliest.

Literary works also included treatises on mathematics, astronomy, medicine, and magic, as well as various religious texts and canonical lists that classified the categories of creation (probably the earliest genre, going back to the beginning of the Old Kingdom, *c.* 2575 BC, or even a little earlier). Among these texts, little is truly systematic, a notable exception being a medical treatise on wounds. The absence of systematic enquiry contrasts with Egyptian practical expertise in such fields as surveying, which was used both for orienting and planning buildings to remarkably fine tolerances and for the regular division of fields after the inundation; the Egyptians also surveyed and established the dimensions of their entire country by the beginning of the Middle Kingdom. These precise tasks required both knowledge of astronomy and highly ingenious techniques, but they apparently were achieved with little theoretical analysis.

Whereas in the earliest periods Egypt seems to have been administered almost as the personal estate of the king, by the central Old Kingdom it was divided into about 35 nomes, or provinces, each with its own officials. Administration was concentrated at the capital, where most of the central elite lived and died. In the nonmonetary Egyptian economy, its essential functions were the collection, storage, and redistribution of produce; the drafting and organization of manpower for specialized labour, probably including irrigation and flood protection works, and major state projects; and the supervision of legal matters. Administration and law were not fully distinct and both depended ultimately on the king. The settlement of disputes was in part an administrative task, for which the chief guiding criterion was precedent, while contractual relations were regulated by the use of standard formulas. State and temple both partook in redistribution and held massive reserves of grain; temples were economic as well as religious institutions. In periods of decentralization similar functions were exercised by local grandees. Markets had only a minor role, and craftsmen were employees who normally traded only what they produced in their free time. The wealthiest officials escaped this pattern to some extent by receiving their income in the form of land and maintaining large establishments that included their own specialized workers.

Administration and law

The essential medium of administration was writing, reinforced by personal authority over the nonliterate 99 percent of the population; texts exhorting the young to be scribes emphasize that the scribe commanded while the rest did the work. Most officials (almost all of whom were men) held several offices and accumulated more as they progressed up a complex ranked hierarchy, at the top of which was the vizier, the chief administrator and judge. The vizier reported to the king, who in theory retained certain powers, such as authority to invoke the death penalty, absolutely.

Before the Middle Kingdom, the civil and the military were not sharply distinguished. Military forces consisted

of local militias under their own officials and included foreigners, and nonmilitary expeditions to extract minerals from the desert or to transport heavy loads through the country were organized in similar fashion. Until the New Kingdom there was no separate priesthood. Holders of civil office also had priestly titles, and priests had civil titles. Often priesthoods were sinecures: their chief significance was the income they brought. The same was true of the minor civil titles accumulated by high officials. At a lower level, minor priesthoods were held on a rotating basis by "laymen" who served every fourth month in temples. State and temple were so closely interconnected that there was no real tension between them before the late New Kingdom.

Sources, calendars, and chronology. For all but the last century of Egyptian prehistory, whose neolithic and later phases are normally termed "predynastic," evidence is exclusively archaeological; later native sources have only mythical allusions to such remote times. The dynastic period of native Egyptian rulers is generally divided into 30 dynasties, following the *Aegyptiaca* of the Greco-Egyptian writer Manetho of Sebennytos (early 3rd century BC), excerpts of which are preserved in later writers. Manetho apparently organized his dynasties by the capital cities from which they ruled, but several of his divisions also reflect political or dynastic changes, that is, changes of the party holding power. He gave the lengths of reign of kings or of entire dynasties and even longer periods. Because of textual corruption and a tendency to inflation, his figures cannot be used to reconstruct chronology and reign lengths without supporting evidence and analysis.

Basis for outline of dynasties

Manetho's prime sources were earlier Egyptian king lists, the organization of which he imitated. The most significant preserved example of a king list is the Turin Canon, a fragmentary papyrus in the Egyptian Museum in Turin, Italy, which originally listed all kings of the 1st through the 17th dynasty, preceded by a mythical dynasty of gods and one of the "spirits, followers of Horus." The document gave reign lengths for individual kings, as well as totals for some dynasties and longer periods.

In early periods the kings' years of reign were not given numbers but were named for salient events, and lists were made of the names. More extensive details were added to the lists for the 4th and 5th dynasties, when dates were assigned according to biennial cattle censuses numbered through each king's reign. Fragments of such lists are preserved on the Palermo Stone, an inscribed piece of basalt (at the Regional Museum of Archaeology in Palermo, Italy), and related pieces in the Cairo Museum and University College London; these are probably all parts of a late copy of an original document.

Calculating dates for Egyptian prehistory

The Egyptians did not date by eras longer than the reign of a single king, so a historical framework must be created from totals of reign lengths, which are then related to astronomical data that may allow whole periods to be fixed precisely. This is done through references to astronomical events and correlations with the three calendars in use in Egyptian antiquity. All dating was by a civil calendar, derived from the lunar calendar, which was introduced in the first half of the 3rd millennium BC. The civil year had 365 days and started in principle when Sirius, or the Dog Star—also known as Sothis (Ancient Egyptian: Il Sopdet)—became visible above the horizon after a period of absence, which at that time occurred some weeks before the Nile began to rise for the inundation. Every four years the civil year advanced one day in relation to the Julian year (with 365¼ days), and after a cycle of about 1,460 years it would again agree with the lunisolar calendar. Religious ceremonies were organized according to two lunar calendars that had months of 29 or 30 days, with extra, intercalary months every three years or so.

Four mentions of the rising of Sirius (generally known as Sothic dates) are preserved in texts from the 3rd to the 1st millennia, but by themselves these references cannot yield an absolute chronology. Such a chronology can be computed from larger numbers of lunar dates and cross-checked from solutions for the observations of Sirius. Various chronologies are in use, however, differing by up to 40 years for the 2nd millennium BC and by more than a century for the beginning of the 1st dynasty. The chronologies offered in most publications up to 1985 have been disproved for the Middle and New kingdoms by a restudy of the evidence for the Sothic and especially the lunar dates. For the 1st millennium, dates in the Third Intermediate Period are approximate; a supposed fixed year of 945 BC, based on links with the Old Testament, turns out to be variable by a number of years. Late Period dates (664–332 BC) are almost completely fixed. Before the 12th dynasty, plausible dates for the 11th can be computed backward, but for earlier times dates are approximate. A total of 955 years for the 1st through the 8th dynasty in the Turin Canon has been used to assign a date of about 3100 BC for the beginning of the 1st dynasty, but this requires excessive average reign lengths, and an estimate of 2925 BC is preferable. Radiocarbon and other scientific dating of samples from Egyptian sites have not improved on, or convincingly contested, computed dates. Recent work on radiocarbon dates from Egypt does, however, yield results encouragingly close to dates computed in the manner described above.

King lists and astronomy give only a chronological framework. A vast range of archaeological and inscriptional sources for Egyptian history survives, but none of it was produced with the interpretation of history in mind. No consistent political history of ancient Egypt can be written. The evidence is very unevenly distributed, there are gaps of many decades, and in the 3rd millennium BC no continuous royal text recording historical events was inscribed. Private biographical inscriptions of all periods from the 5th dynasty (*c.* 2465–*c.* 2325 BC) to the Roman conquest (30 BC) record individual involvement in events but are seldom concerned with their general significance. Royal inscriptions from the 12th dynasty (1938–1756 BC) to Ptolemaic times aim to present a king's actions according to an overall conception of "history," in which he is the re-creator of the order of the world and the guarantor of its continued stability or its expansion. The goal of his action is not to serve humanity but the gods, while nonroyal individuals may relate their own successes to the king in the first instance and sometimes to the gods. Only in the decentralized intermediate periods did the nonroyal recount internal strife. Kings did not mention dissent in their texts unless it came at the beginning of a reign or a phase of action and was quickly and triumphantly overcome in a reaffirmation of order. Such a schema often dominates the factual content of texts, and it creates a strong bias toward recording foreign affairs, because in official ideology there is no internal dissent after the initial turmoil is over. "History" is as much a ritual as a process of events; as a ritual, its protagonists are royal and divine. Only in the Late Period did these conventions weaken significantly. Even then, they were retained in full for temple reliefs, where they kept their vitality into Roman times.

Despite this idealization, the Egyptians were well aware of history, as is clear from their king lists. They divided the past into periods comparable with those used by Egyptologists, and they evaluated the personalities of rulers as the founders of epochs, for salient exploits, or, especially in folklore, for their bad qualities. The Demotic Chronicle, a text of the Ptolemaic period, purports to foretell the bad end that would befall numerous Late Period kings as divine retribution for their wicked actions.

The recovery and study of ancient Egypt. European interest in ancient Egypt was strong in Roman times and revived in the Renaissance, when the small amount of information provided by visitors to the country was compensated for by the wealth of Egyptian remains in the city of Rome. Views of Egypt were dominated by the classical tradition that it was the land of ancient wisdom; this wisdom was thought to inhere in the hieroglyphic script, which was believed to impart profound symbolic ideas, not—as it in fact does—the sounds and words of texts. Between the 15th and 18th centuries, Egypt had a minor but significant position in general views of antiquity, and its monuments gradually became better known through the work of scholars in Europe and travelers in the country itself; the finest publications of the latter were by Richard Pococke, Frederik Ludwig Norden, and Carsten Niebuhr,

all of whose works in the 18th century helped to stimulate an Egyptian revival in European art and architecture. Coptic, the Christian successor of the ancient Egyptian language, was studied from the 17th century, notably by Athanasius Kircher, for its potential to provide the key to Egyptian.

Napoleon's expedition to and short-lived conquest of Egypt in 1798 was the culmination of 18th-century interest in the East. The expedition was accompanied by a team of scholars who recorded the ancient and contemporary country, issuing in 1809–28 the *Description de l'Égypte,* the most comprehensive study to be made before the decipherment of the hieroglyphic script. The Rosetta Stone, which bears a decree of Ptolemy V Epiphanes in hieroglyphs, demotic, and Greek, was discovered during the expedition and was ceded to the British after the French capitulation; it became the property of the British Museum in London. This document greatly assisted the decipherment, accomplished by Jean-François Champollion in 1822.

Discovery of the Rosetta Stone

The Egyptian language revealed by the decipherment and more than 150 years of study is a member of the Afro-Asiatic, or Hamito-Semitic, language family. The Egyptian is closest to the family's Semitic branch but is distinctive in many respects. During several millennia it changed greatly. The script does not write vowels. Because Greek forms for royal names were known from Manetho long before the Egyptian forms became available, those used to this day are a mixture of Greek and Egyptian.

In the first half of the 19th century vast numbers of antiquities were exported from Egypt, forming the nucleus of collections in many major museums. These were removed rather than excavated, inflicting, together with the economic development of the country, colossal damage on ancient sites. At the same time, many travelers and scholars visited the country and recorded the monuments. The most important, and remarkably accurate, record was produced by the Prussian expedition led by Karl Richard Lepsius, in 1842–45, which explored sites as far south as the central Sudan.

Egyptology as a scholarly pursuit

In the mid-19th century Egyptology developed as a subject in France and in Prussia. The Antiquities Service and a museum of Egyptian antiquities were established in Egypt by the French Egyptologist Auguste Mariette, a great excavator who attempted to preserve sites from destruction, and the Prussian Heinrich Brugsch made great progress in the interpretation of texts of many periods and published the first major Egyptian dictionary. In 1880 Flinders (later Sir Flinders) Petrie began more than 40 years of methodical excavation, which created an archaeological framework for all the chief periods of Egyptian culture except for remote prehistory. Petrie was the initiator of much in archaeological method, but he was later surpassed by George Andrew Reisner, who excavated for American institutions from 1899 to 1937. The greatest late 19th-century Egyptologist was Adolf Erman of Berlin, who put the understanding of the Egyptian language on a sound basis and wrote general works that for the first time organized what was known about the earlier periods.

From the 1890s on, complete facsimile copies of Egyptian monuments have been published, providing a separate record that becomes more vital as the originals decay. The pioneer of this epigraphy was Norman de Garis Davies, who was joined in the 1920s by the Oriental Institute of the University of Chicago and other enterprises. Many scholars are now engaged in epigraphy.

In the first half of the 20th century some outstanding archaeological discoveries were made: Howard Carter uncovered the tomb of Tutankhamen in 1922; Pierre Montet found the tombs of 21st–22nd-dynasty kings at Tanis in 1939–44; and W.B. Emery and L.P. Kirwan found tombs of the Ballānah culture (the 4th through the 6th century AD) in Nubia in 1931–34. The last of these was part of the second survey of Lower Nubia in 1929–34, which preceded the second raising of the Aswān Dam. This was followed in the late 1950s and '60s by an international campaign to excavate and record sites in Egyptian and Sudanese Nubia before the completion of the Aswān High Dam in 1970. Lower Nubia is now one of the most thoroughly explored archaeological regions of the world. Most of its many temples have been moved, either to higher ground nearby, as happened to Abu Simbel and Philae, or to quite different places, including various foreign museums. The campaign also had the welcome consequence of introducing a wide range of archaeological expertise to Egypt, so that standards of excavation and recording in the country have risen greatly.

Excavation and survey of great importance continues in many places. For example, at Ṣaqqārah, part of the necropolis of the ancient city of Memphis, new areas of the Sarapeum have been uncovered with rich finds, and a major New Kingdom necropolis is being thoroughly explored. The site of ancient Memphis itself has been systematically surveyed, its position in relation to the ancient course of the Nile has been established, and urban occupation areas have been studied in detail for the first time.

Egyptology is, however, a primarily interpretive subject. There have been outstanding contributions, for example in art, for which Heinrich Schäfer established the principles of the rendering of nature, and in language. New light has been cast on texts, the majority of which are written in a simple metre that can serve as the basis of sophisticated literary works. The physical environment, social structure, kingship, and religion are other fields in which great advances have been made, while the reconstruction of the outline of history is constantly being improved in detail.

THE PREDYNASTIC AND EARLY DYNASTIC PERIODS

Predynastic Egypt. The peoples of predynastic Egypt were the successors of the Paleolithic inhabitants of northeastern Africa, who had spread over much of its area; during wet phases they had left remains in regions as inhospitable as the Great Sand Sea. The final desiccation of the Sahara was not complete until the end of the 3rd millennium BC; over thousands of years people must have migrated from there to the Nile Valley, the environment of which improved as it dried out. In this process, the decisive change from the nomadic hunter-gatherer way of life of Paleolithic times to settled agriculture has not so far been identified. Some time after 5000 BC the raising of crops was introduced, probably on a horticultural scale, in small, local cultures that seem to have penetrated southward through Egypt into the oases and the Sudan. Several of the basic food plants that were grown are native to the Near East, so the new techniques probably spread from there. No large-scale migration need have been involved, and the cultures were at first largely self-contained. The preserved evidence for them is unrepresentative, because it comes from the low desert, where relatively few people lived; as later, most people probably settled in the Valley and Delta.

The earliest known Neolithic cultures in Egypt have been found at Marimda Banī Salāma, on the southwest edge of the Delta, and farther to the southwest, in the Fayyūm. The site at Marimda Banī Salāma, which dates to the 6th–5th millennia BC, gives evidence of settlement and shows that cereals were grown. In the Fayyūm, where evidence dates to the 5th millennium BC, the settlements were near the shore of Lake Qārūn, and the settlers engaged in fishing. Marimda is a very large site that was occupied for many centuries. The inhabitants lived in lightly-built huts; they may have buried their dead within their houses, but areas where burials have been found may not have been occupied by dwellings at the same time. Pottery was used in both cultures. In addition to these Egyptian Neolithic cultures, others have been identified in the Western Desert, in the Second Cataract area, and north of Khartoum. Some of these are as early as the Egyptian ones, while others overlapped with the succeeding Egyptian predynastic cultures.

Tasian and Badarian cultures

In Upper Egypt, between Asyūṭ and Luxor, have been found the Tasian culture (named after Dayr Tasa) and the Badarian culture (named after al-Badārī); these date from the late 5th millennium BC. Most of the evidence for them comes from cemeteries, where the burials included fine black-topped red pottery, ornaments, some copper objects, and glazed steatite beads. The most characteristic predynastic luxury objects, slate palettes for grinding cos-

metics, occur for the first time in this period. The burials show little differentiation of wealth and status and seem to belong to a peasant culture without central political organization.

Probably contemporary with both predynastic and dynastic times are thousands of rock drawings of a wide range of motifs, including boats, found throughout the Eastern Desert, in Lower Nubia, and as far west as Mount ʿUwaynāt, which stands near modern Egypt's borders with Libya and The Sudan in the southwest. The drawings show that nomads were common throughout the desert, probably down to the late 3rd millennium BC, but they cannot be dated precisely; they may all have been produced by nomads, or inhabitants of the Nile Valley may often have penetrated the desert and made drawings.

Naqādah I

Naqādah I, named after the major site of Naqādah but also called Amratian after al-ʿAmīrah, is a distinct phase that succeeded Badarian and has been found as far south as Kawm al-Aḥmar (Hierakonpolis; ancient Egyptian Nekhen), near the sandstone barrier of Jabal al-Silsila, which was the cultural boundary of Egypt in predynastic times. Naqādah I differs from Badarian in its density of settlement and in the typology of its material culture, but hardly at all in the social organization implied by finds. Burials were in shallow pits in which the bodies faced to the west, like those of later Egyptians. Notable types of material found in graves are fine pottery decorated with representational designs in white on red, figurines of men and women, and hard stone mace-heads that are the precursors of important late predynastic objects.

Naqādah II

Naqādah II, also known as Gerzean after al-Girza, is the most important predynastic culture. The heartland of its development was the same as that of Naqādah I, but it spread gradually throughout the country. South of Jabal al-Silsila, sites of the culturally similar Nubian A Group are found as far as the Second Cataract and beyond; these have a long span, continuing as late as the Egyptian Early Dynastic Period. During Naqādah II, large sites developed at Kawm al-Aḥmar, Naqādah, and Abydos, showing by their size the concentration of settlement, as well as exhibiting increasing differentiation in wealth and status. Few sites have been identified between Asyūṭ and the Fayyūm, and this region may have been sparsely settled, perhaps supporting a pastoral rather than agricultural population. Near modern Cairo, at al-ʿUmārī, Maʿādi, and Wādī Digla, and stretching as far south as the latitude of the Fayyūm, are sites of a separate, contemporary culture. Maʿādi was an extensive settlement that traded with the Near East and probably acted as an intermediary for transmitting goods to the south. In this period, imports of lapis lazuli provide evidence that trade networks extended as far afield as Afghanistan.

The material culture of Naqādah II included increasing numbers of prestige objects. The characteristic mortuary pottery is made of buff desert clay, principally from around Qena, and is decorated in red with pictures of uncertain meaning showing boats, animals, and scenes with human figures. Stone vases, many made of hard stones that come from remote areas of the Eastern Desert, are common and of remarkable quality, and cosmetic palettes display elaborate designs, with outlines in the form of animals, birds, or fish. Flint was worked with extraordinary skill to produce large ceremonial knives of a type that continued in use during dynastic times.

Sites of late Naqādah II (sometimes termed Naqādah III) are found throughout Egypt, including the Memphite area and the Delta, and appear to have replaced the local Lower Egyptian cultures. Links with the Near East intensified and some distinctively Mesopotamian motifs and objects were briefly in fashion in Egypt. The cultural unification of the country probably accompanied a political unification, but this must have proceeded in stages and cannot be reconstructed in detail. In an intermediate stage, local states may have formed at Kawm al-Aḥmar, Naqādah, and Abydos, and in the Delta at such sites as Buto (modern Tall al-Farāʿīn) and Sais. Ultimately, Abydos became preeminent; its late predynastic cemetery of Umm al-Qaʿāb was extended to form the burial place of the kings of the 1st dynasty. In the latest predynastic period, objects bearing written symbols of royalty were deposited throughout the country, and primitive writing also appeared in marks on pottery. Because the basic symbol for the king, a falcon on a decorated palace facade, hardly varies, these objects are thought to have belonged to a single line of kings or a single state, and not to a set of small states. This symbol became the royal Horus name, the first element in a king's titulary, which presented the reigning king as the manifestation of an aspect of the god Horus, the leading god of the country. Over the next few centuries several further definitions of the king's presence were added to this one.

Thus at this time Egypt seems to have been a state unified under kings who introduced writing and the first bureaucratic administration. These kings, who could have ruled for more than a century, may correspond with a set of names preserved on the Palermo Stone, but no direct identification can be made between them. The latest was probably Narmer, whose name has been found near Memphis, at Abydos, on a ceremonial palette and mace-head from Kawm al-Aḥmar, and at the Palestinian sites of Tall Gat and ʿArad. The relief scenes on the palette show him wearing the two chief crowns of Egypt and defeating northern enemies, but these probably are stereotyped symbols of the king's power and role and not records of specific events of his reign. They demonstrate that the position of the king in society and its presentation in mixed pictorial and written form had been elaborated by this date.

During this time Egyptian artistic style and conventions were formulated, together with writing. The process led to a complete and remarkably rapid transformation of material culture, so that many dynastic Egyptian prestige objects hardly resemble their forerunners.

The Early Dynastic Period (c. 2925–c. 2575 BC). *The 1st dynasty (c. 2925–c. 2775 BC).* The beginning of the historical period is characterized by the introduction of written records in the form of regnal year names—the records that later were collected in documents such as the Palermo Stone. The first king of Egyptian history, Menes, is therefore a creation of the later record, not the actual unifier of the country; he is known from Egyptian king lists and from classical sources and is credited with irrigation works and with founding the capital, Memphis. On small objects from this time, one of them dated to the important king Narmer but certainly mentioning a different person, there are two possible mentions of a "Men" who may be the king Menes. If these do name Menes, he was probably the same person as Aha, Narmer's probable successor, who was then the founder of the 1st dynasty. Changes in the naming patterns of kings reinforce the assumption that a new dynasty began with his reign. Aha's tomb at Abydos is altogether more grandiose than previously built tombs, while the first of a series of massive tombs at Ṣaqqārah, next to Memphis, supports the tradition that the city was founded then as a new capital. This shift from Abydos is the culmination of intensified settlement in the crucial area between the Valley and the Delta, but Memphis did not yet overcome the traditional pull of its predecessor: the large tombs at Ṣaqqārah appear to belong to high officials, while the kings were buried at Abydos in tombs without formal superstructures. Their mortuary cult may have been conducted in flimsy buildings in designated areas nearer the cultivation, around which a number of burials of important individuals were grouped.

Introduction of written records

In the late predynastic period and the first half of the 1st dynasty, Egypt extended its influence into southern Palestine and probably Sinai and conducted a campaign as far as the Second Cataract. The First Cataract area, with its centre on Elephantine, an island in the Nile opposite the modern town of Aswān, was permanently incorporated into Egypt, but Lower Nubia was not.

Between late predynastic times and the 4th dynasty—and probably early in the period—the Nubian A Group came to an end. There is some evidence that political centralization was in progress around Qustul, but this did not lead to any further development and may indeed have prompted a preemptive strike by Egypt. For Nubia, the malign proximity of the largest state of the time stifled

advancement. During the 1st dynasty, writing spread gradually, but because it was used chiefly for administration, the records, which were kept within the floodplain, have not survived. The artificial writing medium of papyrus was invented by the middle of the 1st dynasty. There was a surge in prosperity, and thousands of tombs of all levels of wealth have been found throughout the country. The richest contained magnificent goods in metal, ivory, and other materials, the most widespread luxury products being extraordinarily fine stone vases. The high point of 1st-dynasty development was the long reign of Den (flourished *c.* 2850 BC).

Elements of the kings' names

During the 1st dynasty three titles were added to the royal Horus name: "Two Ladies," an epithet presenting the king as making manifest an aspect of the protective goddesses of the south (Upper Egypt) and the north (Lower Egypt); "Golden Horus," the precise meaning of which is unknown; and "Dual King," a ranked pairing of the two basic words for king, later associated with Upper and Lower Egypt. These titles were followed by the king's own birth name, which in later centuries was written in a cartouche.

The 2nd dynasty (c. 2775–c. 2650 BC). From the end of the 1st dynasty there is evidence of rival claimants to the throne. One line may have become the 2nd dynasty, whose first king's Horus name, Hetepsekhemwy, means "peaceful in respect of the two powers" and may allude to the conclusion of strife between two factions or parts of the country, to the antagonistic gods Horus and Seth, or to both. Hetepsekhemwy and his successor, Reneb, moved their burial places to Ṣaqqārah; the tomb of the third king, Nynetjer, has not been found. The second half of the dynasty was a time of conflict and rival lines of kings, some of whose names are preserved on stone vases from the 3rd-dynasty Step Pyramid at Ṣaqqārah or in king lists. Among these contenders, Peribsen took the title of Seth instead of Horus and was probably opposed by Horus Khasekhem, whose name is known only from Kawm al-Aḥmar and who used the programmatic epithet "effective sandal against evil." The last ruler of the dynasty combined the Horus and Seth titles to form the Horus-and-Seth Khasekhemwy, "arising in respect of the two powers," to which was added "the two lords are at peace in him." Khasekhemwy was probably the same person as Khasekhem after the successful defeat of his rivals, principally Peribsen. Both Peribsen and Khasekhemwy had tombs at Abydos, and the latter also built a monumental brick funerary enclosure near the main temple (there were two further such enclosures).

The 3rd dynasty (c. 2650–c. 2575 BC). There were links of kinship between Khasekhemwy and the 3rd dynasty, but the change between them is marked by a definitive shift of the royal burial place to Memphis. Its first king, Sanakhte, is attested in reliefs from Maghāra in Sinai. His successor, Djoser (Horus name Netjerykhet), was one of the outstanding kings of Egypt. His Step Pyramid at Ṣaqqārah is both the culmination of an epoch and—as the first large all-stone building, many times larger than anything attempted before—the precursor of later achievements. The pyramid is set in a much larger enclosure than that of Khasekhemwy at Abydos and contains reproductions in stone of ritual structures that had previously been built of perishable materials. Architectural details of columns, cornices, and moldings provided many models for later development. The masonry techniques look to brickwork for models and show little concern for the structural potential of stone. The pyramid itself evolved through numerous stages from a flat mastaba (an oblong tomb with a burial chamber dug beneath it, common at earlier nonroyal sites) into a six-stepped, almost square pyramid. There was a second, symbolic tomb with a flat superstructure on the south side of the enclosure; this probably substituted for the traditional royal burial place of Abydos. The king and some of his family were buried deep under the pyramid, where tens of thousands of stone vases were deposited, a number bearing inscriptions of the first two dynasties. Thus, in perpetuating earlier forms in stone and burying this material, Djoser invoked the past in support of his innovations.

Djoser's Step Pyramid at Ṣaqqārah

Djoser's name was famous in later times and his monument was studied in the Late Period. Imhotep, whose title as a master sculptor is preserved from the Step Pyramid complex, may have been its architect; he lived on into the next reign. His fame also endured, and in the Late Period he was deified and became a god of healing. In Manetho's history he is associated with reforms of writing and this may reflect a genuine tradition, for hieroglyphs were simplified and standardized at this time.

Djoser's successor, Sekhemkhet, planned a still more grandiose step pyramid complex, and a later king, Khaba, began one at Zawyat al-'Aryan, a few miles south of Giza. The burial place of the last king of the dynasty, Huni, is unknown. It has often been suggested that he built the pyramid of Maydūm, but this probably was the work of his successor, Snefru. Inscribed material naming 3rd-dynasty kings is known from Maghāra to Elephantine but not from the Near East or Nubia.

The organizational achievements of the 3rd dynasty are reflected in its principal monument, whose message of centralization and concentration of power is reinforced in a negative sense by the archaeological record. Outside the vicinity of Memphis, the Abydos area continued to be important, and four enormous tombs, probably of high officials, were built at the nearby site of Bayt Khallaf; there were small, nonmortuary step pyramids throughout the country, some of which may date to the 4th dynasty. Otherwise, little evidence comes from the provinces, from which wealth must have flowed to the centre, leaving no rich local elite. By the 3rd dynasty the rigid structure of the later nomes, or provinces, which formed the basis of Old Kingdom administration, had been created, and the imposition of its uniform pattern may have impoverished local centres. Tombs of the elite at Ṣaqqārah, notably those of Hezyre and Khabausokar, contained artistic masterpieces that look forward to the Old Kingdom.

THE OLD KINGDOM (c. 2575–c. 2130 BC) AND THE FIRST INTERMEDIATE PERIOD (c. 2130–1938 BC)

The Old Kingdom. *The 4th dynasty (c. 2575–c. 2465 BC).* The first king of the 4th dynasty, Snefru, probably built the step pyramid of Maydūm and then modified it to form the first true pyramid. Due west of Maydūm was the small step pyramid of Saylah, in the Fayyūm, at which Snefru also worked. He built two pyramids at Dahshūr; the southern of the two is known as the Bent Pyramid because its upper part has a shallower angle of inclination than its lower part. This difference may be due to structural problems or may have been planned from the start, in which case the resulting profile may reproduce a solar symbol of creation. The northern Dahshūr pyramid, the later of the two, has the same angle of inclination as the upper part of the Bent Pyramid and a base area exceeded only by that of the Great Pyramid. Both pyramids had mortuary complexes attached to them. Snefru's building achievements were thus at least as great as those of any later king and introduced a century of unparalleled construction.

Snefru's monuments

In a long perspective, the 4th dynasty was an isolated phenomenon, a period when the potential of centralization was realized to its utmost and a disproportionate amount of the state's resources was used on the kings' mortuary provisions, almost certainly at the expense of general living standards. No significant 4th-dynasty sites have been found away from the Memphite area. Tomb inscriptions show that high officials were granted estates scattered over many nomes, especially in the Delta. This pattern of landholding may have avoided the formation of local centres of influence while encouraging intensive exploitation of the land. People who worked on these estates were not free to move, and they paid a high proportion of their earnings in dues and taxes. The building enterprises must have relied on drafting vast numbers of men, probably after the harvest had been gathered in the early summer and during part of the inundation.

Snefru's was the first king's name that was regularly written inside the cartouche, an elongated oval that is one of the most characteristic Egyptian symbols. The cartouche itself is older and was shown as a gift bestowed by gods on

the king, signifying long duration on the throne. It soon acquired associations with the sun, so that its first use by the builder of the first true pyramid, which is probably also a solar symbol, is not coincidental.

The Great Pyramid at Giza

Snefru's successor, Khufu (Cheops), built the Great Pyramid at Giza, to which were added the slightly smaller second pyramid of one of Khufu's sons, Khafre (more correctly Rekhaef, the Chephren of Greek sources), and that of Menkaure (Mycerinus). Khufu's successor, his son Redjedef, began a pyramid at Abū Ruwaysh, and a king of uncertain name began one at Zawyat al-'Aryan. The last known king of the dynasty (there was probably one further), Shepseskaf, built a monumental mastaba at south Ṣaqqārah and was the only Old Kingdom ruler not to begin a pyramid. These works, especially the Great Pyramid, show a great mastery of monumental stoneworking: individual blocks were large or colossal and were very accurately fitted to one another. Surveying and planning also were carried out with remarkable precision.

Apart from the colossal conception of the pyramids themselves, the temple complexes attached to them show great mastery of architectural forms. Khufu's temple or approach causeway was decorated with impressive reliefs, fragments of which were incorporated in the 12th-dynasty pyramid of Amenemhet I at al-Lisht. The best known of all Egyptian sculpture, Khafre's Great Sphinx at Giza and his extraordinary seated statue of Nubian gneiss, date from the middle 4th dynasty.

The Giza pyramids form a group of more or less completed monuments surrounded by many tombs of the royal family and the elite, hierarchically organized and laid out in neat patterns. This arrangement contrasts with that of the reign of Snefru, when important tombs were built at Maydūm and Ṣaqqārah, while the King was probably buried at Dahshūr. Of the Giza tombs, only those of the highest-ranking officials were decorated: except among the immediate entourage of the kings, the freedom of expression of officials was greatly restricted. Most of the highest officials were members of the very large royal family, so that power was concentrated by kinship as well as other means. This did not prevent factional strife: the complex of Redjedef was deliberately and thoroughly destroyed, probably at the instigation of his successor, Khafre.

Egyptian expansion into Nubia

The Palermo Stone records a campaign to Lower Nubia in the reign of Snefru that may be associated with graffiti in the area itself. The Egyptians founded a settlement at Buhen, at the north end of the Second Cataract, which endured for 200 years; others may have been founded between there and Elephantine. The purposes of this penetration were probably to establish trade farther south and to create a buffer zone. No archaeological traces of a settled population in Lower Nubia have been found for the Old Kingdom period: the oppressive presence of Egypt seems to have robbed the inhabitants of their resources, rather as the Egyptian provinces were exploited in favour of the king and the elite.

Snefru and the builders of the Giza pyramids represented a classic age to later times. Snefru was the prototype of a good king, whereas Khufu and Khafre had tyrannical reputations, perhaps only because of the size of their monuments. Little direct evidence for political or other attitudes survives from the dynasty, in part because writing was only just beginning to be used for recording continuous texts. Many great works of art were, however, produced for kings and members of the elite, and these set a pattern for later work. Kings of the 4th dynasty identified themselves, at least from the time of Redjedef, as Son of Re (the sun god); worship of the sun god reached a peak in the 5th dynasty.

The 5th dynasty (c. 2465–c. 2325 BC). The first two kings of the 5th dynasty, Userkaf and Sahure, were sons of a lady, Khentkaues, who was a member of the 4th-dynasty royal family. The third king, Neferirkare, may also have been her son. A story from the Middle Kingdom that makes them all sons of a priest of Re may derive from a tradition that they were true worshipers of the sun god and implies, probably falsely, that the 4th-dynasty kings were not. Six kings of the 5th dynasty displayed their devotion to the sun god by building personal temples to his cult. These temples, of which the two so far identified are sited similarly to pyramids, probably had a mortuary significance for the king as well as honouring the god. The kings' pyramids should therefore be seen in conjunction with the temples, some of which received lavish endowments and were served by many high-ranking officials.

Pyramids have been identified for seven of the nine kings of the dynasty, at Ṣaqqārah (Userkaf and Unas, the last king), Abū Ṣīr (Sahure, Neferirkare, Reneferef, and Nuserre), and south Ṣaqqārah (Djedkare Izezi, the eighth king). The pyramids are smaller and less solidly constructed than those of the 4th dynasty, but the reliefs from their mortuary temples are better preserved and of very fine quality; that of Sahure gives a fair impression of their decorative program. The interiors contained religious scenes relating to provision for Sahure in the next life, while the exteriors presented his "historical" role and relations with the gods. Sea expeditions to Lebanon to acquire timber are depicted, as are aggression against and capture of Libyans. Despite their apparent precision, in which captives are named and total figures given, these scenes may not refer to specific events, for the same motifs with the same details were frequently shown over the next 250 years; Sahure's use of them might not have been the earliest.

Relations with foreign lands

Foreign connections were far-flung. Goldwork of the period has been found in Anatolia, while stone vases named for Khafre and Pepi I (6th dynasty) have been found at Tall Mardīkh in Syria, the capital of the important state of Ebla, which was destroyed around 2250 BC. The absence of 5th-dynasty evidence from the site is probably a matter of chance. Expeditions to the turquoise mines of Sinai continued as before. In Nubia, graffiti and inscribed seals from Buhen document Egyptian presence until late in the dynasty, when control was probably abandoned in the face of immigration from the south and the deserts; later generations of the immigrants are known as the Nubian C Group. From the reign of Sahure on, there are records of trade with Punt, a partly legendary land probably in the region of Eritrea, from which the Egyptians obtained incense and myrrh, as well as exotic African products that had been traded from still farther afield. Thus the reduced level of royal display in Egypt does not imply a less prominent general role for the country.

Organization of the country's administration

High officials of the 5th dynasty were no longer members of the royal family, although a few married princesses. Their offices still depended on the king, and in their biographical inscriptions they presented their exploits as relating to him, but they justified other aspects of their social role in terms of a more general morality. They progressed through their careers by acquiring titles in complex ranked sequences that were manipulated by kings throughout the 5th and 6th dynasties. This institutionalization of officialdom has an archaeological parallel in the distribution of elite tombs, which no longer clustered so closely around pyramids. Many are at Giza, but the largest and finest are at Ṣaqqārah and Abū Ṣīr. The repertory of decorated scenes in them continually expanded, but there was no fundamental change in their subject matter. Toward the end of the 5th dynasty some officials with strong local ties began to build their tombs in the Nile Valley and the Delta, in a development that symbolized the elite's slowly growing independence from royal control.

Something of the working of the central administration is visible in papyri from the mortuary temples of Neferirkare and Reneferef at Abū Ṣīr. These show well-developed methods of accounting and meticulous recordkeeping and document the complicated redistribution of goods and materials between the royal residence, the temples, and officials who held priesthoods. Despite this evidence for detailed organization, the consumption of papyrus was modest and cannot be compared, for example, with that of Greco-Roman times.

The last three kings of the dynasty, Menkauhor, Djedkare Izezi, and Unas, did not have personal names compounded with "-Re," the name of the sun god (Djedkare is a name assumed on accession); and Izezi and Unas did not build solar temples. Thus there was a slight shift away from the solar cult. The shift could be linked with the rise

of Osiris, the god of the dead, who is first attested from the reign of Neuserre. His origin was, however, probably some centuries earlier. The pyramid of Unas, whose approach causeway was richly decorated with historical and religious scenes, is inscribed inside with spells intended to aid the deceased in the hereafter; varying selections of the spells occur in all later Old Kingdom pyramids. (As a collection they are known as the Pyramid Texts.) Many of the spells were old when they were inscribed; their presence documents the increasing use of writing rather than a change in beliefs. The Pyramid Texts show the importance of Osiris, at least for the king's passage into the next world: it was an undertaking that aroused anxiety and had to be assisted by elaborate rituals and spells.

The 6th dynasty (c. 2325–c. 2150 BC). No marked change can be discerned between the reigns of Unas and Teti, the first king of the 6th dynasty. Around Teti's pyramid in the northern portion of Ṣaqqārah was built a cemetery of large tombs, including those of several viziers. Together with tombs near the pyramid of Unas, this is the latest group of private monuments of the Old Kingdom in the Memphite area.

Information on 6th-dynasty political and external affairs is more abundant because inscriptions of high officials were longer. Whether the circumstances they describe were also typical of less loquacious ages is unknown, but the very existence of such inscriptions is evidence of a tendency to greater independence among officials. One, Weni, who lived from the reign of Teti through those of Pepi I and Merenre, was a special judge in the trial of a conspiracy in the royal household, mounted several campaigns against a region east of Egypt or in southern Palestine, and organized two quarrying expeditions. In the absence of a standing army, the Egyptian force was levied from the provinces by officials from local administrative centres and other settlements; there were also contingents from several southern countries and a tribe of the Eastern Desert.

Trading expeditions in the 6th dynasty

Three biographies of officials from Elephantine record trading expeditions to the south in the reigns of Pepi I and Pepi II. The location of the regions named in them is debated and may have been as far afield as the Butāna, south of the Fifth Cataract. Some of the trade routes ran through the Western Desert, where the Egyptians established an administrative post at Balāṭ in ad-Dākhilah Oasis, some distance west of al-Khārijah Oasis. Egypt no longer controlled Lower Nubia, which was settled by the C Group and formed into political units of gradually increasing size, possibly as far as Karmah, south of the Third Cataract; relations with this state deteriorated into armed conflict in the reign of Pepi II. Karmah was the southern cultural successor of the Nubian A Group and became an urban centre in the late 3rd millennium BC, remaining Egypt's chief southern neighbour for seven centuries. To the north, the Karmah state stretched as far as the Second Cataract and at times farther still. Its southern extent has not been determined, but sites of similar material culture are scattered over vast areas of the central Sudan.

Increase in provincialization

The provincializing tendencies of the late 5th dynasty continued in the 6th, especially during the extremely long reign (up to 94 years) of Pepi II. Increasing numbers of officials resided in the provinces, amassed local offices, and emphasized local concerns, including religious leadership, in their inscriptions. At the capital the size and splendour of the cemeteries decreased, and some tombs of the end of the dynasty were decorated only in their subterranean parts, as if security could not be guaranteed aboveground. The pyramid complex of Pepi II at southern Ṣaqqārah, which was probably completed in the first 30 years of his reign, stands out against this background as the last major monument of the Old Kingdom, comparable with its predecessors in artistic achievement. Three of his queens were buried in small pyramids around his own; these are the only known queens' monuments inscribed with Pyramid Texts.

The 7th and 8th dynasties (c. 2150–30 BC). Pepi II was followed by several ephemeral rulers, who were in turn succeeded by the short-lived 7th dynasty of Manetho's history (from which no king's name is known) and the 8th, one of whose kings, Ibi, built a small pyramid at southern Ṣaqqārah. Several 8th-dynasty kings are known from inscriptions found in the temple of Min at Qift in the south; this suggests that their rule was recognized throughout the country. The instability of the throne is, however, a sign of political decay, and the fiction of centralized rule may have been accepted only because there was no alternative style of government to kingship.

Collapse of the Old Kingdom state

With the end of the 8th dynasty the Old Kingdom state collapsed. About this time there was widespread famine and violence; the consequent rise in the death rate can be seen in sharply increased numbers of burials in cemeteries. The country emerged impoverished and decentralized from this episode, the prime cause of which may have been political failure, environmental disaster, or, more probably, a combination of the two. In this period the desiccation of northeastern Africa reached a peak, producing conditions similar to those of modern times, and a related succession of low inundations may have coincided with the decay of central political authority. These environmental changes are, however, only approximately dated and their relationship with the collapse cannot be proved.

The First Intermediate Period. *The 9th dynasty (c. 2130–2080 BC).* After the end of the 8th dynasty the throne passed to kings from Heracleopolis, who made their native city the capital, although Memphis continued to be important. They were acknowledged throughout the country, but inscriptions of nomarchs (chief officials of nomes) in the south show that the kings' rule was nominal. At Dara, north of Asyūṭ, for example, a local ruler called Khety styled himself king and built a pyramid with a surrounding "courtly" cemetery. At al-Mi'alla, south of Luxor, Ankhtify, the nomarch of the al-Jabalayn region, recorded his annexation of the Idfū nome and extensive raiding in the Theban area. Ankhtify acknowledged an unidentifiable king Neferkare but campaigned with his own troops. Major themes of inscriptions of the period are the nomarch's provision of food supplies for his people in times of famine and his success in promoting irrigation works. Artificial irrigation had probably long been practiced, but exceptional poverty and crop failure made concern with it worth recording. Inscriptions of Nubian mercenaries employed by local rulers in the south indicate how entrenched military action was.

The 10th (c. 2080–c. 1970 BC) and 11th (2081–1938 BC) dynasties. A period of generalized conflict focused on twin dynasties at Thebes and Heracleopolis. The latter, the 10th, probably continued the line of the 9th. The founder of the 9th or 10th dynasty was named Khety and the dynasty as a whole was termed the House of Khety. Several Heracleopolitan kings were named Khety; another important name is Merikare. Whereas the Theban dynasty was stable, kings succeeded one another rapidly at Heracleopolis. There was continual conflict, and the boundary between the two realms shifted around the region of Abydos. As yet, the course of events in this period cannot be reconstructed.

Several major literary texts purport to describe the upheavals of the First Intermediate Period, the "Instruction for Merikare," for example, being ascribed to one of the kings of the 9th or 10th dynasty. These texts led earlier Egyptologists to posit a Heracleopolitan literary flowering, but there is now a tendency to date them to the Middle Kingdom, so that they would have been written with enough hindsight to allow a more effective critique of the sacred order. The "Heracleopolitan Age" may therefore be a fiction.

Until the 11th dynasty made Thebes its capital, Hermonthis (modern Armant), on the west bank of the Nile, had been the centre of the Theban nome. The dynasty honoured as its ancestor the God's Father Mentuhotep, probably the father of its first king, Inyotef I (2081–65 BC), whose successors were Inyotef II and Inyotef III (2065–16 and 2016–08 BC, respectively). The fourth king, Mentuhotep I (sometimes numbered II; 2008–1957 BC, whose throne name was Nebhepetre), gradually reunited Egypt and ousted the Heracleopolitans, changing his titulary in stages to record his conquests. Around his 20th regnal year he assumed the Horus name Divine of the White Crown,

implicitly claiming all of Upper Egypt. By his regnal year 42 this was changed to Uniter of the Two Lands, a traditional royal epithet that he revived with a literal meaning and presented in a new, emphatic iconography. In later times Mentuhotep was celebrated as the founder of the epoch now known as the Middle Kingdom. His remarkable mortuary complex at Dayr al-Baḥrī, which seems to have had no pyramid, was the architectural inspiration for Hatshepsut's later structure built alongside.

In the First Intermediate Period, monuments were set up by a slightly larger section of the population and, in the absence of central control, internal dissent and conflicts of authority became visible in public records. Nonroyal individuals took over some of the privileges of royalty, notably identification with Osiris in the hereafter and the use of the Pyramid Texts; these were incorporated into a more extensive corpus inscribed on coffins (and hence termed the Coffin Texts) and continued to be inscribed during the Middle Kingdom. The unified state of the Middle Kingdom did not reject these acquisitions and so had a broader cultural basis than the Old Kingdom.

THE MIDDLE KINGDOM (1938–c. 1600 BC) AND THE SECOND INTERMEDIATE PERIOD (c. 1630–1540 BC)

The Middle Kingdom. Mentuhotep I campaigned in Lower Nubia, where he may have been preceded by the Inyotefs. In Thebes he built a novel and impressive mortuary complex at Dayr al-Baḥrī, which served as inspiration for Hatshepsut's adjacent temple 500 years later. The complex contained some of the earliest known depictions of Amon-Re, the dynastic god of the Middle Kingdom and the New Kingdom. Mentuhotep I was himself deified and worshiped, notably in the Aswān area. In administration, he attempted to break the power of the nomarchs, but his policy was unsuccessful in the longer term.

Mentuhotep I's mortuary complex at Dayr al-Baḥrī

Mentuhotep I's successors, Mentuhotep II (1957–45 BC) and Mentuhotep III (1945–38 BC) also ruled from Thebes. The reign of Mentuhotep III corresponds to seven years marked "missing" in the Turin Canon, and he may later have been deemed illegitimate. Records of a quarrying expedition to the Wadi Ḥammāmāt (Wādī Rawḍ ʿĀyd) from his second regnal year were inscribed on the order of his vizier Amenemhet, who almost certainly usurped the throne and founded the 12th dynasty. Not all the country welcomed the 11th dynasty, the monuments and self-presentation of which remained local and Theban.

The 12th dynasty (1938–c. 1756 BC). In a text probably circulated as propaganda during the reign of Amenemhet I (1938–08 BC), the time preceding his reign is depicted as a period of chaos and despair, from which a saviour called Ameny from the extreme south was to emerge. This presentation may well be stereotyped, but there could have been armed struggle before he seized the throne. Nonetheless, his mortuary complex at al-Lisht contained monuments on which his name was associated with that of his predecessor. In style, his pyramid and mortuary temple looked back to Pepi II of the end of the Old Kingdom, but the pyramid was built of mud brick with a stone casing and consequently is badly ruined.

Amenemhet I moved the capital back to the Memphite area, founding a residence named Itj-towy "[Amenemhet is] he who takes possession of the Two Lands," which was for later times the archetypal royal residence. Itj-towy was probably situated between Memphis and the pyramids of Amenemhet I and Sesostris I (at modern al-Lisht), while Memphis remained the centre of population. From later in the dynasty there is the earliest evidence for a royal palace (not a capital) in the eastern Delta. The return to the Memphite area was accompanied by a revival of Old Kingdom artistic styles, in a resumption of central traditions that contrasted with the local ones of the 11th dynasty. In his policy toward the nomarchs, Amenemhet retreated from the absolutism of the Mentuhoteps, and major tombs of the first half of the dynasty, which display considerable local independence, are preserved at several sites, notably Beni Hasan, Mayr, and Qau. After the second reign of the dynasty, no more important private tombs were constructed at Thebes, but several kings made benefactions to Theban temples.

In his 20th regnal year, Amenemhet I took his son Sesostris I (or Senwosret, 1918–1875 BC) as his co-regent, presumably in order to avoid the instability of the First Intermediate Period and its aftermath. This practice was followed in the next two reigns and recurred sporadically in later times. During the following 10 years of joint rule Sesostris undertook campaigns in Lower Nubia that led to its conquest as far as the central area of the Second Cataract. A series of fortresses was begun in the region and there was a full occupation, but the local C Group population was not integrated culturally with the conquerors.

Sesostris I's campaigns to the south

Amenemhet I apparently was murdered during Sesostris' absence on a campaign to Libya, but Sesostris was able to maintain his hold on the throne without major disorder. He consolidated his father's achievements, but, in one of the earliest preserved inscriptions recounting royal exploits, he spoke of internal unrest. An inscription of the next reign alludes to campaigns to Syria-Palestine in the time of Sesostris; whether these were raiding expeditions and parades of strength, in what was then a seminomadic region, or whether a conquest was intended or achieved, is not known. It is clear, however, that the traditional view that the Middle Kingdom hardly intervened in the Near East is incorrect.

In the early 12th dynasty the written language was regularized in its classical form of Middle Egyptian, a rather artificial idiom that was probably always somewhat removed from the vernacular. The first datable corpus of literary texts was composed in Middle Egyptian. Two of these relate directly to political affairs and offer fictional justifications for the rule of Amenemhet I and Sesostris I, respectively. Several that are ascribed to Old Kingdom authors or that describe events of the First Intermediate Period, but are composed in Middle Egyptian, probably also date from around this time. The most significant of these is the "Instruction for Merikare," a discourse on kingship and moral responsibility. It is often used as a source for the history of the First Intermediate Period but may preserve no more than a memory of its events. Most of these texts continued to be copied in the New Kingdom.

Little is known of the reigns of Amenemhet II (1876–42 BC) and Sesostris II (1844–37 BC). These kings built their pyramids in the Fayyūm, while also beginning an intensive exploitation of its agricultural potential that reached a peak in the reign of Amenemhet III (1818–1770 BC). The king of the 12th dynasty with the most enduring reputation was Sesostris III (1836–18 BC), who extended Egyptian conquests to Semna, at the south end of the Second Cataract, while also mounting at least one campaign to Palestine. Sesostris III completed an extensive chain of fortresses in the Second Cataract; at Semna he was worshiped as a god in the New Kingdom.

Frequent campaigns and military occupation, which lasted another 150 years, required a standing army. A force of this type may have been created early in the 12th dynasty but becomes better attested near the end. It was based on "soldiers," whose title means literally "citizens," levied by district, and officers of several grades and types. It was separate from New Kingdom military organization and seems not to have enjoyed very high status.

Egypt's standing army

The purpose of the occupation of Lower Nubia is disputed, because the size of the fortresses and the level of manpower needed to occupy them might seem disproportionate to local threats. An inscription of Sesostris III set up in the fortresses emphasizes the weakness of the Nubian enemy, while a boundary marker and fragmentary papyri show that the system channeled trade with the south through the central fortress of Mirgissa. The greatest period of the Karmah state to the south was still to come, but for centuries it had probably controlled a vast stretch of territory. The best explanation of the Egyptian presence is that Lower Nubia was annexed by Egypt whenever possible, while Karmah was a rival worth respecting and preempting; in addition, the physical scale of the fortresses may have become something of an end in itself. It is not known whether Egypt wished similarly to conquer Palestine, but an inscription of Sesostris' reign records a campaign in Palestine, and numerous administrative seals of the period have been found there.

Sesostris III's administrative reorganization

Sesostris III finally broke the power of the nomarchs and reorganized Egypt into four regions corresponding to the northern and southern halves of the Nile Valley and the eastern and western Delta. Rich evidence for middle-ranking officials from the religious centre of Abydos, and for administrative practice in documents from al-Lahūn, conveys an impression of a pervasive, centralized bureaucracy, which later came to run the country under its own momentum. The prosperity created by peace, conquests, and agricultural development is visible in royal monuments and monuments belonging to the minor elite, but there was no small, powerful, and wealthy group of the sort seen in the Old and New Kingdoms. Sesostris III and his successor, Amenemhet III (1818–*c.* 1770 BC), left a striking artistic legacy in the form of statuary depicting them as aging, careworn rulers, probably alluding to a conception of the suffering king known from literature of the dynasty. This departure from the bland ideal, which may have sought to bridge the gap between king and subjects in the aftermath of the attack on elite power, was not taken up in later times.

The reigns of Amenemhet III and Amenemhet IV (*c.* 1770–1760 BC) and Sebeknefru (*c.* 1760–1756 BC), the first certainly attested female monarch, were apparently peaceful, but the accession of a woman marked the end of the dynastic line.

The 13th dynasty (c. 1756–c. 1630 BC). Despite a continuity of outward forms and of the rhetoric of inscriptions between the 12th and 13th dynasties, there was a complete change in kingship. In little more than a century about 70 kings occupied the throne. Many can have reigned only for months, and there were probably rival claimants to the throne, but in principle the royal residence remained at Itj-towy and the kings ruled the whole country. Egypt's hold on Lower Nubia was maintained, as was its position as the leading state in the Near East. Large numbers of private monuments document the prosperity of the official classes, and a proliferation of titles is evidence of their continued expansion. In government the vizier assumed prime importance, and a single family held the office for much of a century.

Waves of immigration

Asiatic immigration is known in the late 12th dynasty and became widespread in the 13th. From the late 18th century BC the northeastern Delta was settled by successive waves of Palestinians, who retained their own material culture. Starting with the "Instruction for Merikare," Egyptian texts warn against the dangers of infiltration of this sort, and its occurrence shows a weakening of government. There may also have been a rival dynasty, called the 14th, at Xois in the north central Delta, but this is known only from Manetho's history and could have had no more than local significance. Several late 13th-dynasty kings are attested only at Thebes and may have formed a rival line or moved their residence there from the north. Toward the end of this period Egypt lost control of Lower Nubia, where the garrisons, which had been regularly replaced with fresh troops, settled and were partly assimilated. The Karmah state overran and incorporated the region. Some Egyptian officials resident in the Second Cataract area served the new rulers. The site of Karmah has yielded many Egyptian artifacts, including old pieces pillaged from their original contexts. Most were items of trade between the two countries, some probably destined for exchange against goods imported from sub-Saharan Africa. Around the end of the Middle Kingdom and during the Second Intermediate Period, Medjay tribesmen from the Eastern Desert settled in the Nile Valley from around Memphis to the Third Cataract. Their presence is marked by distinctive shallow graves with black-topped pottery, and they have traditionally been termed the "Pan-grave" culture by archaeologists. They were assimilated culturally in the New Kingdom, but the word Medjay came to mean police or militia; they probably came as mercenaries.

The Second Intermediate Period. The increasing competition for power in Egypt and Nubia crystallized in the formation of two new dynasties: the 15th, called the Hyksos (*c.* 1630–*c.* 1523 BC), with its capital at Avaris (Tall ad-Dab'a) in the Delta, and the 17th (*c.* 1630–1540 BC), ruling from Thebes. The word Hyksos goes back to an Egyptian phrase meaning "ruler of foreign lands" and occurs in Manetho's narrative cited in the works of the Jewish historian Josephus (1st century AD), which depicts the new rulers as sacrilegious invaders who despoiled the land. They may have invaded, but they presented themselves—with the exception of the title Hyksos—as Egyptian kings and appear to have been accepted as such. The main line of Hyksos was acknowledged throughout Egypt and may have been recognized as overlords in Palestine, but they tolerated other lines of kings, both those of the 17th dynasty and the various minor Hyksos who are termed the 16th dynasty. The 17th dynasty therefore had to accept that it was a junior line, and in this distinction of status lay an occasion, if not a cause, of later conflict. The 15th dynasty consisted of six kings, the best known being the fifth, Apopis, who reigned for up to 40 years. There were many 17th-dynasty kings, probably belonging to several different families. The northern frontier of the Theban domain was at al-Qusiyya, but there was trade across the border and the Thebans pastured their herds in the Delta.

Rule of the Hyksos

Asiatic rule brought many technical innovations to Egypt, as well as cultural innovations such as new musical instruments and musical styles. The changes affected techniques from bronze working and pottery to looms; and new breeds of animals and new crops were introduced. In warfare, composite bows, new types of daggers and scimitars, and above all the horse and chariot transformed previous practice, although the chariot may ultimately have been as important as a prestige vehicle as for tactical advantages it conferred. The effect of these changes was to bring Egypt, which had been technologically backward, onto the level of western Asia. Because of these advances and the perspectives it opened up, Hyksos rule was decisive for Egypt's later empire in the Near East.

Whereas the 13th dynasty was fairly prosperous, the Second Intermediate Period may have been impoverished. The regional centre of the cult of Osiris at Abydos, which has produced the largest quantity of Middle Kingdom monuments, lost importance, but sites such as Thebes, Idfū, and Kawm al-Aḥmar have yielded significant, if sometimes crudely worked, remains. Virtually no information has come from the north, where the Hyksos ruled, and it is impossible to assess their impact on the economy or on high culture. The Second Intermediate Period was the consequence of political fragmentation and immigration and was not associated with the severe economic collapse of the early First Intermediate Period.

Kamose's challenge of the Hyksos

Toward the end of the 17th dynasty (*c.* 1545 BC), the Theban king Seqenenre challenged Apopis, probably dying in battle against him. Seqenenre's successor, Kamose, renewed the challenge, stating in an inscription that it was intolerable to share his land with an Asiatic and a Nubian (the Karmah ruler). By the end of his third regnal year he had made raids as far south as the Second Cataract (and possibly much farther) and in the north to the neighbourhood of Avaris, also intercepting in the Western Desert a letter sent from Apopis to a new Karmah ruler on his accession. By campaigning to the north and to the south Kamose acted out his implicit claim to the territory ruled by Egypt in the Middle Kingdom. His exploits formed a vital stage in the long struggle to expel the Hyksos.

(J.R.Ba.)

THE NEW KINGDOM

The 18th dynasty. *Ahmose.* Although Ahmose (ruled *c.* 1539–14 BC) had been preceded by Kamose, who was either his father or brother, Egyptian tradition regarded Ahmose as the founder of a new dynasty because he was the native ruler who reunified Egypt. Continuing a recently inaugurated practice, he married his full sister Ahmose-Nofretari. The queen was given the title of God's Wife of Amon. Like her predecessors of the 17th dynasty, Queen Ahmose-Nofretari was influential and highly honoured. A measure of her importance was her posthumous veneration at Thebes, where later pharaohs were depicted offering to her as a goddess among the gods.

Ahmose was very young at his accession, and his campaigns to expel the Hyksos from the Delta and regain former Egyptian territory to the south probably started around

his 10th regnal year. Destroying the Hyksos stronghold at Avaris, in the eastern Delta, he finally drove them beyond the eastern frontier and then besieged Sharuhen (Tall al-Farʿah) in southern Palestine; the full extent of his conquests may have been much greater. His penetration of the Near East came at a time when there was no major established power in the region. This political gap facilitated the creation of an Egyptian "empire."

Ahmose's officers and soldiers were rewarded with spoil and captives, who became personal slaves. This marked the creation of an influential military class. Like Kamose, Ahmose campaigned as far south as Buhen. For the administration of the regained territory he created a new office, overseer of southern foreign lands, which ranked second only to the vizier. Its incumbent was accorded the honorific title of king's son, indicating that he was directly responsible to the king as deputy.

Administration

The early New Kingdom bureaucracy was modeled after that of the Middle Kingdom. The vizier was the chief administrator and the highest judge of the realm. By the middle of the 15th century BC the office had been divided into two, one vizier for Upper and one for Lower Egypt. During the 18th dynasty some young bureaucrats were educated in temple schools, reinforcing the integration of civil and priestly sectors. Early in the dynasty many administrative posts were inherited, but royal appointment of capable officials, often selected from military officers who had served the king on his campaigns, later became the rule. The trend was thus away from bureaucratic families and the inheritance of office.

Amenhotep I. Ahmose's son and successor, Amenhotep I (ruled *c.* 1514–1493 BC), pushed the Egyptian frontier southward to the Third Cataract, near the capital of the Karmah state, while also gathering tribute from his Asiatic possessions and perhaps campaigning in Syria. The emerging kingdom of Mitanni in northern Syria, which is first mentioned on a stela of one of Amenhotep's soldiers and was also known by the name of Nahrin, may have threatened Egypt's conquests to the north.

The New Kingdom saw increased devotion to the state god Amon-Re, whose cult gave the king, as his representative, the mission of expanding Egypt's frontiers. Amon-Re benefited as Egypt was enriched by the spoils of war. Riches were turned over to the god's treasuries, and the king had sacred monuments constructed at Thebes. Under Amenhotep I the pyramidal form of royal tomb was abandoned in favour of a rock-cut tomb, and, except for Akhenaton, all subsequent New Kingdom rulers were buried in concealed tombs in the famous Valley of the Kings in western Thebes. Separated from the tombs, royal mortuary temples were erected at the edge of the desert. Perhaps because of this innovation, Amenhotep I later became the patron deity of the workmen who excavated and decorated the royal tombs. The location of his own tomb is unknown.

Thutmose I and Thutmose II. Lacking a surviving heir, Amenhotep I was succeeded by one of his generals, Thutmose I (ruled 1493–*c.* 1482 BC), who married his own full sister Ahmose. In the south Thutmose destroyed the Karmah state. He inscribed a rock as a boundary marker, later confirmed by Thutmose III, near Kanisa-Kurgus, north of the Fifth Cataract. He then executed a brilliant campaign into Syria and across the Euphrates, where he erected a victory stela near Carchemish.

Thus in the reign of Thutmose I, Egyptian conquests in the Near East and Africa reached their greatest extent, but they may not yet have been firmly held. His little-known successor, Thutmose II (*c.* 1482–79 BC), continued his policies.

Hatshepsut and Thutmose III. At Thutmose II's death his queen and sister, Hatshepsut, had only a young daughter; but a minor wife had borne him a boy, who served as a priest in the Temple of Amon. This son, Thutmose III (ruled 1479–26 BC), later reconquered Egypt's Asiatic empire and became an outstanding ruler. During his first few regnal years Thutmose III theoretically controlled the land, but Hatshepsut governed as regent. Sometime between Thutmose III's second and seventh regnal years she assumed the kingship herself. According to one version of the event, the oracle of Amon proclaimed her king at Karnak, where she was crowned. A more propagandistic account, preserved in texts and reliefs of her splendid mortuary temple at Dayr al-Baḥrī, ignores the reign of Thutmose II and asserts that her father, Thutmose I, proclaimed her as his successor. Upon becoming king, Hatshepsut became the dominant partner in a joint rule that lasted until her death in about 1458 BC; there are monuments dedicated by Hatshepsut that depict both kings. She had the support of various powerful personalities, who did not, however, form a homogeneous faction; the most notable among them was Senenmut, the steward and tutor of her daughter Neferure. In styling herself king, Hatshepsut adopted the royal titulary but avoided the epithet "mighty bull," regularly employed by other kings. Although in her reliefs she was depicted as a male, pronominal references in the texts generally reflect her womanhood. Similarly, much of her statuary shows her in male form, but there are rarer examples that render her as a woman. In less formal documents she was referred to as "King's Great Wife," that is, "Queen," while Thutmose III was "King." There is thus a certain ambiguity in the treatment of Hatshepsut as king.

Hatshepsut's assumption of the kingship

Her temple reliefs depict pacific enterprises, such as the transporting of obelisks for Amon's temple and a commercial expedition to Punt; her art style looked back to Middle Kingdom ideals. Some warlike scenes are depicted, however, and she may have waged a campaign in Nubia. In one inscription she blamed the Hyksos for the supposedly poor state of the land before her rule, even though they had been expelled from the region more than a generation earlier.

During Hatshepsut's ascendancy Egypt's position in Asia deteriorated because of the expansion of Mitannian power in Syria. Shortly after her death, the Prince of Kadesh, a Syrian city, stood with troops of 330 princes of a Syro-Palestinian coalition at Megiddo; such a force was more than merely defensive and the intention may have been to advance against Egypt. The 330 must have represented all the places of any size in the region that were not subject to Egyptian rule and may be a schematic figure derived from a list of place-names. It is noteworthy that Mitanni itself was not directly involved.

Thutmose III proceeded to Gaza with his army and then to Yehem, subjugating rebellious Palestinian towns along the way. His annals relate how, at a consultation concerning the best route over the Mount Carmel ridge, the King overruled his officers and selected a shorter but more dangerous route through the ʿArūnah Pass and then led the troops himself. The march went smoothly, and when the Egyptians attacked at dawn they prevailed over the enemy troops and besieged Megiddo.

Thutmose III meanwhile coordinated the landing of other army divisions on the Syro-Palestinian littoral, whence they proceeded inland, so that the strategy resembled a pincer technique. The siege ended in a treaty by which Syrian princes swore an oath of submission to the King. As was normal in ancient diplomacy and in Egyptian practice, the oath was binding only upon those who swore it, not upon future generations.

By the end of the first campaign Egyptian domination extended northward to a line linking Byblos and Damascus. Although the Prince of Kadesh remained to be vanquished, Assyria sent lapis lazuli as tribute; Asiatic princes surrendered their weapons, including a large number of horses and chariots. Thutmose III took only a limited number of captives. He appointed Asiatic princes to govern the towns and took their brothers and sons to Egypt, where they were educated at the court. Most eventually returned home to serve as loyal vassals, though some remained in Egypt at court. In order to ensure the loyalty of Asiatic city-states, Egypt maintained garrisons that could quell insurrection and supervise the delivery of tribute. There never was an elaborate Egyptian imperial administration in Asia.

Thutmose III conducted numerous subsequent campaigns in Asia. The submission of Kadesh was finally achieved, but Thutmose III's ultimate aim was the defeat of Mitanni. He used the navy to transport troops to

Asiatic coastal towns, avoiding arduous overland marches from Egypt. His great eighth campaign led him across the Euphrates; although the countryside around Carchemish was ravaged, the city was not taken, and the Mitannian prince was able to flee. The psychological gain of this campaign was perhaps greater than its military success, for Babylonia, Assyria, and the Hittites all sent tribute in recognition of Egyptian dominance. Although Thutmose III never subjugated Mitanni, he placed Egypt's conquests on a firm footing by constant campaigning that contrasts with the forays of his predecessors. His annals inscribed in the temple of Karnak are remarkably succinct and accurate, but his other texts, notably one set up in his newly founded Nubian capital of Napata, are more conventional in their rhetoric.

Thutmose III initiated a truly imperial Egyptian rule in Nubia. Much of the land became estates of institutions in Egypt, while local cultural traits disappear from the archaeological record. Sons of chiefs were educated at the Egyptian court; a few returned to Nubia to serve as administrators—and some were buried there in Egyptian fashion. Nubian fortresses lost their strategic value and became administrative centres. Open towns developed around them, and in several temples outside their walls the cult of the divine king was established. Lower Nubia supplied gold from the desert and hard and semiprecious stones. From farther south came African woods, perfumes, oil, ivory, panther skins, and ostrich plumes. There is scarcely any trace of local population from the later New Kingdom, when many more temples were built in Nubia; by the end of the 20th dynasty the region had almost no prosperous settled population.

Under Thutmose III the wealth of empire became apparent in Egypt. Many temples were built and vast sums were donated to the estate of Amon-Re. There are many tombs of his high officials at Thebes. The capital had been moved to Memphis, but Thebes remained the religious centre.

The campaigns of kings like Thutmose III required a large military establishment, including a hierarchy of officers and a very expensive chariotry. The king grew up with military companions whose close connection with him enabled them to participate increasingly in government. Military officers were appointed to high civil and religious positions, and by the Ramesside period the influence of such people came to outweigh that of the traditional bureaucracy.

Amenhotep II. About two years before his death Thutmose III appointed his 18-year-old son, Amenhotep II (ruled *c.* 1426–1400 BC), as co-regent. Just prior to his father's death, Amenhotep II set out on a campaign to an area near Kadesh, in Syria, whose city-states were now caught up in the power struggle between Egypt and Mitanni; Amenhotep II killed seven princes and shipped their bodies back to Egypt to be suspended from the ramparts of Thebes and Napata. In his seventh and ninth years Amenhotep II made further campaigns into Asia, where the Mitannian king pursued a more vigorous policy. The revolt of the important coastal city of Ugarit was a serious matter, because Egyptian control over Syria required bases along the littoral for inland operations and the provisioning of the army. Ugarit was pacified, and the fealty of Syrian cities, including Kadesh, was reconfirmed.

Amenhotep II's campaigns into Asia

Thutmose IV. Amenhotep II's son Thutmose IV (ruled 1400–1390 BC) sought to establish peaceful relations with the Mitannian king Artatama, who had been successful against the Hittites. Artatama gave his daughter in marriage, the prerequisite for which was probably the Egyptian cession of some Syrian city-states to the Mitannian sphere of influence. This was the first such diplomatic marriage, paving the way for the age of Amenhotep III, when the emphasis shifted from war to diplomacy and the enjoyment of the luxury of empire.

Foreign influences during the early 18th dynasty. During the empire period Egypt maintained commercial ties with Phoenicia, Crete, and the Aegean islands. The Egyptians portrayed goods obtained through trade as foreign tribute. In the Theban tombs there are representations of Syrians bearing Aegean products and of Aegeans carrying Syrian bowls and amphorae—indicative of close commercial interconnections among Mediterranean lands. Egyptian ships trading with Phoenicia and Syria journeyed beyond to Crete and the Aegean, a route that explains the occasional confusion of products and ethnic types in Egyptian representations. The most prized raw material from the Aegean world was silver, which was lacking in Egypt, where gold was relatively abundant.

One result of empire was a new appreciation of foreign culture. Not only were foreign objets d'art imported into Egypt but Egyptian artisans imitated Aegean wares as well. Imported textiles inspired the ceiling patterns of Theban tomb chapels, and Aegean art with its spiral motifs and rendition of movement influenced Egyptian artists. Under Amenhotep II, Asiatic gods are found in Egypt: Astarte and Resheph became revered for their reputed potency in warfare, and Astarte was honoured also in connection with medicine, love, and fertility. Some Asiatic gods were eventually identified with similar Egyptian deities; thus, Astarte was associated with Sekhmet, the goddess of pestilence, and Resheph with Mont, the war god. Just as Asiatics resident in Egypt were incorporated into Egyptian society and could rise to important positions, so their gods, though represented as foreign, were worshiped according to Egyptian cult practices. The breakdown of Egyptian isolationism and increased cosmopolitanism in religion are also reflected in hymns that praise Amon-Re's concern for the welfare of Asiatics.

Amenhotep III. Thutmose IV's son Amenhotep III (ruled 1390–53 BC) acceded to the throne at about the age of 12. He soon wed Tiy, who became his queen. Earlier in the dynasty military men had served as royal tutors; but Tiy's father was a commander of the chariotry, and through this link the royal line became even more directly influenced by the military. In his fifth year Amenhotep III claimed a victory over Cushite rebels, but the Viceroy of Cush, the southern portion of Nubia, probably actually led the troops. The campaign may have led into the Butana, west of the ʿAṭbarah River, farther south than any previous Egyptian military expedition had gone. Several temples erected under Amenhotep III in Upper Nubia between the Second and Third cataracts attest to the importance of the region.

Peaceful relations prevailed with Asia, where control of Egypt's vassals was successfully maintained. A commemorative scarab from the king's 10th year announced the arrival in Egypt of the Mitannian princess Gilukhepa, along with 317 women; thus, another diplomatic marriage helped maintain friendly relations between Egypt and its former foe. Another Mitannian princess was later received into Amenhotep III's harem, and during his final illness the Hurrian goddess Ishtar of Nineveh was sent to his aid. At the expense of older bureaucratic families and the principle of inheritance of office, military men acquired high posts in the civil administration. Most influential was the aged scribe and commander of the elite troops, Amenhotep, son of Hapu, whose reputation as a sage survived into the Ptolemaic period.

Amenhotep III sponsored building on a colossal scale, especially in the Theban area. At Karnak he erected the huge third pylon, and at Luxor he dedicated a magnificent new temple to Amon. The King's own mortuary temple in western Thebes was unrivaled in its size; little remains of it today, but its famous Colossi of Memnon testify to its proportions. He also built a huge harbour and palace complex nearby. Some colossal statues served as objects of public veneration, before which men could appeal to the king's *ka,* which represented the transcendent aspect of kingship. In Karnak, statues of Amenhotep, son of Hapu, were placed to act as intermediaries between supplicants and the gods.

Among the highest-ranking officials at Thebes were men of Lower Egyptian background, who constructed large tombs with highly refined decoration. An eclectic quality is visible in the tombs, certain scenes of which were inspired by Old Kingdom reliefs. The revolutionary art of the succeeding Amarna period perhaps reflects a reaction against the studied perfection of Theban art. The earliest preserved important New Kingdom monuments from Memphis also date from this reign. Antiquarianism is ev-

idenced in Amenhotep III's celebration of his *sed* festivals (rituals of renewal celebrated after 30 years of rule), which were performed at his Theban palace in accordance, it was claimed, with ancient writings. Tiy, whose role was much more prominent than that of earlier queens, participated in these ceremonies.

Amenhotep III's last years were spent in ill health. To judge from his mummy and less formal representations of him from Amarna, he was obese when, in his 38th regnal year, he died and was succeeded by his son Amenhotep IV (ruled 1353–36 BC), the most controversial of all the kings of Egypt.

Amenhotep IV (Akhenaton). The earliest monuments of Amenhotep IV, who in his fifth regnal year changed his name to Akhenaton ("one useful to Aton"), are conventional in their iconography and style, but from the first he gave the sun god a didactic title naming Aton, the solar disk. This title was later written inside a pair of cartouches, as a king's name would be. The king declared his religious allegiance by the unprecedented use of "high priest of the sun god" as one of his own titles. The term Aton had long been in use, but under Thutmose IV the Aton had been referred to as a god, and under Amenhotep III those references became more frequent. Thus, Akhenaton did not create a new god but rather singled out this aspect of the sun god from among others. He also carried further radical tendencies that had recently developed in solar religion, in which the sun god was freed from his traditional mythological context and presented as the sole beneficent provider for the entire world. The King's own divinity was emphasized: the Aton was said to be his father, of whom he alone had knowledge, and they shared the status of king and celebrated jubilees together.

In his first five regnal years, Akhenaton built many temples to the Aton, of which the most important were in the precinct of the temple of Amon-Re at Karnak. In these open-air structures was developed a new, highly stylized form of relief and sculpture in the round. The Aton was depicted not in anthropomorphic form but as a solar disk from which radiating arms extend the hieroglyph for "life" to the noses of the king and his family. During the construction of these temples the cult of Amon and other gods was suspended, and the worship of the Aton in an open-air sanctuary superseded that of Amon, who had dwelt in a dark shrine of the Karnak temple. The King's wife Nefertiti, whom he had married before his accession, was prominent in the reliefs and had a complete shrine dedicated to her that included no images of the King. Her prestige continued to grow for much of the reign.

The transfer of the capital to Amarna

At about the time that he altered his name to conform with the new religion, the King transferred the capital to a virgin site at Amarna (now Tell el-Amarna) in Middle Egypt. There, he constructed a well-planned city—Akhetaton ("The Horizon of Aton")—comprising temples to the Aton, palaces, official buildings, villas for the high ranking, and extensive residential quarters. In the eastern desert cliffs surrounding the city, tombs were excavated for the courtiers; and deep within a secluded wadi the royal sepulchre was prepared. Reliefs in these tombs have been invaluable for reconstructing life at Amarna. The tomb reliefs and stelae portray the life of the royal family with an unprecedented degree of intimacy. They also show that the city was laid out as a great stage, on which the king's daily journeys from palace to city and back made manifest the passage of the sun across the sky.

In Akhenaton's ninth year a more monotheistic didactic name was given to the Aton, and an intense persecution of the older gods, especially Amon, was undertaken. Amon's name was excised from many older monuments throughout the land, and occasionally the word "gods" was expunged. This evidence suggests that the King's monotheistic fervour intensified.

Akhenaton's religious and cultural revolution was highly personal. The peculiar depiction of his physiognomy became the norm for representing not only members of the royal family but commoners as well. In religion the accent was upon the sun's life-sustaining power, and naturalistic scenes adorned the walls and even the floors of Amarna buildings. The king's role in determining the composition of the court is expressed in epithets given to officials he selected from the lesser ranks of society, including the military. Few officials had any connection with the old ruling elite, and some courtiers who had been accepted at the beginning of the reign were purged. Even at Amarna the new religion was not widely accepted below the level of the elite; numerous small objects relating to traditional beliefs have been found at the site.

Akhenaton's revolutionary intent is visible in all of his actions. In representational art, many existing conventions that had no special religious meaning were reversed to emphasize the break with the past. Such a procedure is comprehensible because traditional values were consistently incorporated in cultural expression as a whole; in order to change one part it was necessary to change the whole.

A vital innovation was the introduction of current vernacular forms into the written language. This led in later decades to the creation of new styles for monumental inscriptions and for everyday use. The latter variant, which is now known as Late Egyptian, was not fully developed until the later 19th dynasty.

Akhenaton's violent changes could not have been accomplished without the military, who are ubiquitous in the reliefs, especially from the Karnak temples. Akhenaton's foreign policy and use of force abroad are less well understood. He mounted one minor campaign in Nubia. In the Near East, Egypt's hold on its possessions was not as secure as earlier, but the cuneiform tablets found at Amarna recording his diplomacy are difficult to interpret because the vassals who requested aid from him exaggerated their plight. One reason for unrest in the region was the decline of Mitanni and the resurgence of the Hittites. Between the reign of Akhenaton and the end of the 18th dynasty, Egypt lost control of much territory in Syria.

The aftermath of Amarna. Akhenaton had six daughters by Nefertiti and one or two sons, perhaps by a secondary wife Kiya or by his own daughter Maketaton, who may have died in childbirth and whose infant son is shown in the royal tomb at Amarna. His immediate, ephemeral successor was a woman, possibly his eldest daughter Meritaton. Either she or the widow of Tutankhamen called on the Hittite king Suppiluliumas to supply a consort because she could find none in Egypt; a prince Zannanza was sent, but he was murdered as he reached Egypt. Thus Egypt never had a diplomatic marriage in which a foreign man was received into the country.

Succession of Tutankhaten (Tutankhamen)

After the brief rule of Smenkhkare (1335–32 BC), possibly a son of Akhenaton, Tutankhaten, a nine-year-old child, succeeded and was married to the much older Ankhesenpaaten, Akhenaton's third daughter. Around his third regnal year, the King moved his capital to Memphis, abandoned the Aton cult, and changed his and the Queen's names to Tutankhamen and Ankhesenamen. In an inscription recording Tutankhamen's actions for the gods, the Amarna period is described as one of misery and of the withdrawal of the gods from Egypt. This change, made in the name of the young king, was probably the work of high officials. The most influential were Ay, known by the title God's Father, who served as vizier and regent (his title indicates a close relationship to the royal family), and the general Horemheb, who functioned as royal deputy and whose tomb at Ṣaqqārah contains remarkable scenes of Asiatic captives being presented to the King.

Just as Akhenaton had adapted and transformed the religious thinking that was current in his time, the reaction to the religion of Amarna was influenced by the rejected doctrine. In the new doctrine, all gods were in essence three: Amon, Re, and Ptah (to whom Seth was later added), and in some ultimate sense they too were one. The earliest evidence of this triad is on a trumpet of Tutankhamen and is related to the naming of the three chief army divisions after these gods; religious and secular life were not separate. This concentration on a small number of essential deities may possibly be related to the piety of the succeeding Ramesside period, because both viewed the cosmos as being thoroughly permeated with the divine.

Under Tutankhamen a considerable amount of building was accomplished in Thebes. His Luxor colonnade bears detailed reliefs of the traditional beautiful festival of Opet;

at Karnak he decorated a structure with warlike scenes. He affirmed his legitimacy by referring back to Amenhotep III, whom he called his father. Tutankhamen's modern fame comes from the discovery of his rich burial in the Valley of the Kings. His tomb equipment was superior in quality to the fragments known from other royal burials, and the opulent display—of varying aesthetic value—represents Egyptian wealth at the peak of the country's power.

Ay and Horemheb. Tutankhamen's funeral in about 1323 BC was conducted by his successor, the aged Ay (ruled 1323–19 BC), who in turn was succeeded by Horemheb. The latter probably ruled from 1319 to *c.* 1292 BC, but the length of his poorly attested reign is not certain. Horemheb dismantled many monuments erected by Akhenaton and his successors and used the blocks as fill for huge pylons at Karnak. In this process Nefertiti's image seems to have been defaced more than others. At Karnak and Luxor he appropriated Tutankhamen's reliefs by surcharging the latter's cartouches with his own. Horemheb appointed new officials and priests not from established families but from the army. His policies concentrated on domestic problems. He issued police regulations dealing with the misbehaviour of palace officials and personnel, and he reformed the judicial system, reorganizing the courts and selecting new judges.

The Ramesside period (19th and 20th dynasties). Horemheb was the first post-Amarna king to be considered legitimate in the 19th dynasty, which looked to him as the founder of an epoch. Having no son, he selected his general and vizier, Ramses, to succeed him.

Ramses I and Seti I. Ramses I (ruled 1292–90 BC) hailed from the eastern Delta, and with the 19th dynasty there was a political shift into the Delta. Ramses I was succeeded by his son and co-regent, Seti I, who buried his father and provided him with mortuary buildings at Thebes and Abydos.

Seti I (ruled 1290–79 BC) was a successful military leader who reasserted authority over Egypt's weakened empire in the Near East. The Mitanni state had been dismembered and the Hittites had become the dominant Asiatic power. Before tackling them, Seti laid the groundwork for military operations in Syria by fighting farther south against nomads and Palestinian city-states; then, following the strategy of Thutmose III, he secured the coastal cities and gained Kadesh. Although his engagement with the Hittites was successful, Egypt acquired only temporary control of part of the north Syrian plain. A treaty was concluded with the Hittites who, however, subsequently pushed farther southward and regained Kadesh by the time of Ramses II. Seti I ended a new threat to Egyptain security when he defeated Libyans attempting to enter the Delta. He also mounted a southern campaign, probably to the Fifth Cataract region.

Seti I's reign looked for its model to the mid-18th dynasty and was a time of considerable prosperity. Seti I restored countless monuments that had been defaced in the Amarna period, and the refined decoration of his monuments, particularly his temple at Abydos, shows a classicizing tendency. He also commissioned striking and novel reliefs showing stages of his campaigns, which are preserved notably on the north wall of the great hypostyle hall at Karnak. This diversity of artistic approach is characteristic of the Ramesside period, which was culturally and ethnically pluralistic.

Ramses II. Well before his death, Seti I appointed his son Ramses II, sometimes called Ramses the Great, as crown prince. During the long reign of Ramses II (1279–13 BC) there was a prodigious amount of building, ranging from religious edifices throughout Egypt and Nubia to a new cosmopolitan capital, Pi Ramesse (Tall ad-Dab'a), in the eastern Delta; his cartouches were carved ubiquitously, often on earlier monuments. Ramses II's penchant for decorating vast temple walls with battle scenes gives the impression of a mighty warrior king. His campaigns were, however, relatively few, and after the first decade his reign was peaceful. The most famous scenes record the battle of Kadesh, fought in his fifth regnal year. These and extensive accompanying texts present the battle as an Egyptian victory, but in fact the opposing Hittite coalition fared at least as well as the Egyptians. After this inconclusive struggle, his officers advised him to make peace, saying, "There is no reproach in reconciliation when you make it." In succeeding years Ramses II campaigned in Syria; after a decade of stalemate, a treaty in his 21st year was concluded with Hattusilis III, the Hittite king.

Treaty with the Hittites

The rise of Assyria and unrest in western Anatolia encouraged the Hittites to accept this treaty, while Ramses II may have feared a new Libyan threat to the western Delta. Egyptian and Hittite versions of the treaty survive. It contained a renunciation of further hostilities, a mutual alliance against outside attack and internal rebellion, and the extradition of fugitives. The gods of both lands were invoked as witnesses. The treaty was further cemented 13 years later by Ramses II's marriage to a Hittite princess.

The King had an immense family by his numerous wives, among whom he especially honoured Nefertari. He dedicated a temple to her at Abu Simbel, in Nubia, and built a magnificent tomb for her in the Valley of the Queens.

For the first time in more than a millennium, princes were prominently represented on the monuments. Ramses II's fourth surviving son, Khaemwese, was famous as high priest of Ptah at Memphis. He restored many monuments in the Memphite area, including pyramids and pyramid temples of the Old Kingdom, and had buildings constructed near the Sarapeum at Ṣaqqārah. He was celebrated into Roman times as a sage and magician and became the hero of a cycle of stories.

Merneptah. Ramses II's 13th son, Merneptah (ruled 1213–04 BC), was his successor. Several of Merneptah's inscriptions, of unusual literary style, treat an invasion of the western Delta in his fifth year by Libyans, supported by groups of Sea Peoples who had traveled from Anatolia to Libya in search of new homes. The Egyptians defeated this confederation and settled captives in military camps to serve as Egyptian mercenaries.

One of the inscriptions concludes with a poem of victory (written about another battle), famous for its words, "Israel is desolated and has no seed." This is the earliest documented mention of Israel; it is generally assumed that the exodus of the Jews from Egypt took place under Ramses II.

Merneptah was able to hold most of Egypt's possessions, although early in his reign he had to reassert Egyptian suzerainty in Palestine, destroying Gezer in the process. Peaceful relations with the Hittites and respect for the treaty of Ramses II are indicated by Merneptah's dispatch of grain to them during a famine and by Egyptian military aid in the protection of Hittite possessions in Syria.

Last years of the 19th dynasty. Upon the death of Merneptah, competing factions within the royal family contended for the succession. Merneptah's son Seti II (ruled 1204–1198 BC) had to face a usurper, Amenmeses, who rebelled in Nubia and was accepted in Upper Egypt. His successor, Siptah, was installed on the throne by a Syrian royal butler, Bay, who had become chancellor of Egypt. Siptah was succeeded by Seti II's widow Tausert, who ruled as king from 1193 to 1190 BC, counting her regnal years from the death of Seti II, whose name she restored over that of Siptah. A description in a later papyrus of the end of the dynasty alludes to a Syrian usurper, probably Bay, who subjected the land to harsh taxation and treated the gods as mortals with no offerings in their temples.

The early 20th dynasty: Setnakht and Ramses III. Order was restored by a man of obscure origin, Setnakht (ruled 1190–87 BC), the founder of the 20th dynasty, who appropriated Tausert's tomb in the Valley of the Kings. An inscription of Setnakht recounts his struggle to pacify the land, which ended in the second of his three regnal years.

Setnakht's son Ramses III (ruled 1187–56 BC) was the last great king of the New Kingdom. There are problems in evaluating his achievements because he emulated Ramses II and copied numerous scenes and texts of Ramses II in his mortuary temple at Madīnat Habu, one of the best preserved temples of the empire period. Thus, the historicity of certain Nubian and Syrian wars depicted as his accomplishments is subject to doubt. He did, however, fight battles that were more decisive than any fought by

Ramses II. In his fifth year Ramses III defeated a large-scale Libyan invasion of the Delta in a battle in which thousands of the enemy perished.

A greater menace lay to the north, where a confederation of Sea Peoples was progressing by land and sea toward Egypt. This alliance of obscure tribes came south in the aftermath of the destruction of the Hittite empire. In his eighth regnal year Ramses III engaged them successfully on two frontiers—a land battle in Palestine and a naval engagement in one of the mouths of the Delta. Because of these two victories, Egypt did not undergo the political turmoil or experience the rapid technical advance of the early Iron Age in the Near East. Forced away from the borders of Egypt, the Sea Peoples sailed farther westward, and some of their groups may have given their names to the Sicilians, Sardinians, and Etruscans. The Philistine and Tjekker peoples, who had come by land, were established by the Egyptians in military camps in the southern Palestinian coastal district in an area where the overland trade route to Syria was threatened by attacks by nomads. Initially settled to protect Egyptian interests, these groups later became independent of Egypt. Ramses III used some of these peoples as mercenaries, even in battle against their own kinfolk. In his 11th year he successfully repulsed another great Libyan invasion by the Meshwesh tribes. Meshwesh prisoners of war, branded with the king's name, were settled in military camps in Egypt, and in later centuries their descendants became politically important because of their ethnic cohesiveness and their military role.

These great defensive wars drained the Egyptian economy. Under Ramses III the estate of Amon received only one-fifth as much gold as in Thutmose III's time. Although there are artistic masterpieces at Madīnat Habu, much of the relief inside the temple and the quality of the masonry betray a decline. Toward the end of his reign, administrative inefficiency and the deteriorating economic situation resulted in the government's failure to deliver grain rations on time to necropolis workers, whose dissatisfaction was expressed in demonstrations and in the first recorded strikes in history. Such demonstrations continued sporadically throughout the dynasty. A different sort of internal trouble originated in the royal harem, where a minor queen plotted unsuccessfully to murder Ramses III so that her son might become king. Involved in the plot were palace and harem personnel, government officials, and army officers. A special court of 12 judges was formed to try the accused, who received the death sentence.

Harem conspiracy against Ramses III

Many literary works date to the Ramesside period. Earlier works in Middle Egyptian were copied in schools and in good papyrus copies, and new texts were composed in Late Egyptian. Notable among the latter are stories, several with mythological or allegorical content, that look to folk models rather than to the elaborate written literary types of the Middle Kingdom.

Ramses IV. Ramses III was succeeded by his son Ramses IV (ruled 1156–50 BC). In an act of piety that also reinforced his legitimacy, Ramses IV saw to the compilation of a long papyrus in which the deceased Ramses III confirmed the temple holdings throughout Egypt; Ramses III had provided the largest benefactions to the Theban temples, in terms of donations of both land and personnel. Most of these probably endorsed earlier donations, to which each king added his own gifts. Of the annual income to temples, 86 percent of the silver and 62 percent of the grain was awarded to Amon. The document demonstrates the economic power of the Theban temples, for the tremendous landholdings of Amon's estate throughout Egypt involved the labour of a considerable portion of the population; but the ratio of temple to state income is not known, and the two were not administratively separate. In addition, the temple of Amon, which figures prominently in the papyrus, included within its estates the King's own mortuary temple, for Ramses III was himself deified as a form of Amon-Re, known as Imbued with Eternity.

The later Ramesside kings. The Ramesside period saw a tendency toward the formation of high-priestly families, which kings sometimes tried to counter by appointing outside men to the high priesthood. One such family had developed at Thebes in the second half of the 19th dynasty, and Ramses IV tried to control it by installing Ramessesnakht, the son of a royal steward, as Theban high priest. Ramessesnakht participated in administrative as well as priestly affairs; he personally led an expedition to the Wadi Ḥammāmāt (modern Wādī Rawḍ ʿĀ'id) quarries in the Eastern Desert, and at Thebes he supervised the distribution of rations to the workmen decorating the royal tomb. Under Ramses V (ruled 1150–45 BC), Ramessesnakht's son not only served as steward of Amon but also held the post of administrator of royal lands and chief taxing master. Thus, this family acquired extensive authority over the wealth of Amon and over state finances; but to what extent this threatened royal authority is uncertain. Part of the problem in evaluating the evidence is that Ramesside history is viewed from a Theban bias, because Thebes is the major source of information. Evidence from Lower Egypt, where the king normally resided, is meagre because of unfavourable conditions there for the preservation of monuments or papyri.

Rise in power of high-priestly families

A long papyrus from the reign of Ramses V contains valuable information on the ownership of land and taxation. In Ramesside Egypt most of the land belonged to the state and the temples, while most peasants served as tenant farmers. Some scholars interpret this document as indicating that the state retained its right to tax temple property, at an estimated one-tenth of the crop.

Ramses VI (ruled 1145–37 BC), probably a son of Ramses III, usurped much of his two predecessors' work, including the tomb of Ramses V; a papyrus refers to a possible civil war at Thebes. Following the death of Ramses III the Asiatic empire had rapidly withered away, and Ramses VI is the last king whose name appears at the Sinai turquoise mines. The next two Ramses (ruled 1137–26 BC) were obscure rulers, whose sequence has been questioned. During the reigns of Ramses IX (ruled 1126–08 BC) and Ramses X (1108–04 BC) there are frequent references in the papyri to the disruptions of marauding Libyans near the Theban necropolis.

By the time of Ramses IX the Theban high priest had attained great local influence, though he was still outranked by the king. Early in the reign of Ramses XI (ruled 1104–*c.* 1075 BC), during a civil war at Thebes, the high priest Amenhotep, the son of Ramessesnakht, was suppressed from his office for nine months; the King called upon Panehsy, the viceroy of Cush, to restore order and the fighting spread as far north as Middle Egypt. By Ramses XI's 19th regnal year the Viceroy was driven back and the new high priest of Amon, Herihor—who seems to have had a military background and also claimed the vizierate and the office of Viceroy of Cush—controlled the Theban area. In reliefs at the temple of Khons at Karnak, Herihor was represented as high priest of Amon in scenes adjoining those of Ramses XI. This in itself was unusual, but subsequently he took an even bolder step in having himself depicted as king to the exclusion of the still-reigning Ramses XI. Herihor's kingship was restricted to Thebes, where these years were referred to as a "repeating of [royal] manifestations," which lasted a decade.

With the shrinkage of the empire, the supply of silver and copper was cut off, and the amount of gold entering the economy was reduced considerably. During the reign of Ramses IX the economically distressed inhabitants of western Thebes were found to have pillaged the tombs of kings and nobles (already a common practice in the latter case); the despoiling continued into the reign of Ramses XI, and even the royal mortuary temples were stripped of their valuable furnishings. Nubian troops, called in to restore order at Thebes, themselves contributed to the depredation of monuments. This pillaging brought fresh gold and silver into the economy, and the price of copper rose. The price of grain, which had been inflating, dropped.

Despoiling of tombs

While Ramses XI was still king, Herihor died and was succeeded as high priest by Piankh, a man of similar military background. A series of letters from Thebes tell of Piankh's military venture in Nubia against the former Viceroy of Cush, while Egypt was on the verge of losing control of the south. With the death of Ramses XI, the governor of Tanis, Smendes, became king, founding the 21st dynasty (known as the Tanite).

The Ramesside growth of priestly power was matched by increasingly overt religiosity. Private tombs, the decoration of which had been mostly secular, came to include only religious scenes; oracles were invoked in many kinds of decisions; and private letters contain frequent references to prayer and to regular visits to small temples to perform rituals or consult oracles. The common expression used in letters, "I am all right today; tomorrow is in the hands of god," reflects the ethos of the age. This fatalism, which emphasizes that the god may be capricious and that his wishes cannot be known, is also typical of late New Kingdom Instruction Texts, which show a marked change from their Middle Kingdom forerunners by moving toward a passivity and quietism that suits a less expensive age.

Some of the religious material of the Ramesside period exhibits changes in conventions of display, and some categories have no parallel in the less abundant earlier record, but the shift is real as well as apparent. In its later periods, Egyptian society, the values of which had previously tended to be centralized, secular, and political, became more locally based and more thoroughly pervaded by religion, looking to the temple as the chief institution.

EGYPT FROM 1075 BC TO THE MACEDONIAN INVASION

The Third Intermediate Period (1075–656 BC). *The 21st dynasty.* At the end of the New Kingdom, then, Egypt was divided. The north was inherited by the Tanite 21st dynasty (1075–*c.* 950 BC), and much of the Nile Valley came under the control of the Theban priests (the northern frontier of their domain was the fortress town of al-Hība). Some Theban priests locally assumed the title of king, but there is no indication of conflict between the priests and the Tanite pharaohs. Indeed, the dating of documents, even at Thebes, was in terms of the Tanite reigns, and apparently there were close family ties between the pharaohs and the Thebans. Piankh's son, Pinudjem I, who relinquished the office of high priest and assumed the kingship at Thebes, was probably the father of the Tanite pharaoh Psusennes I. Some members of both the Theban priestly and the Tanite royal lines had Libyan names. With the coming of the new dynasty, and possibly a little earlier, the Meshwesh Libyan military elite, which had been settled mainly in the north by Ramses III, penetrated the ruling group, although it did not become dominant until the 22nd dynasty.

Beginning with Herihor and continuing through the 21st dynasty, the high priests' activities included the pious rewrapping and reburial of New Kingdom royal mummies. The ransacking of the royal tombs during the 20th dynasty necessitated the transfer of the royal remains in stages to two caches—the tomb of Amenhotep II and a cliff tomb at Dayr al-Baḥrī—where they remained undisturbed until modern times. Dockets pertaining to the reburial of these mummies contain important chronological data from the 21st dynasty.

The burials of King Psusennes I (ruled *c.* 1045–*c.* 997 BC) and his successor, Amenemope (ruled *c.* 998–*c.* 989 BC), were discovered at Tanis, but little is known of their reigns. This was a period of the usurpation of statuary and the reuse of material of earlier periods. At Karnak, Pinudjem I, who decorated the facade of the Khons temple, usurped a colossal statue of Ramses II, and Psusennes I's splendid sarcophagus from Tanis had originally been carved for Merneptah. Much of the remains from Tanis comprises material transported from other sites, notably from Pi Ramesse.

After the demise of Egypt's Asiatic empire, the kingdom of Israel eventually developed under the kings David and Solomon. During David's reign, Philistia served as a buffer between Egypt and Israel; but upon David's death the next to the last king of the 21st dynasty, Siamon, invaded Philistia and captured Gezer. If Egypt had any intention of attacking Israel, Solomon's power forestalled Siamon, who presented Gezer to Israel as a dowry in the diplomatic marriage of his daughter to Solomon. This is indicative of the reversal of Egypt's status in foreign affairs since the time of Amenhotep III, who had written the Babylonian king, "From of old, a daughter of the king of Egypt has not been given to anyone."

Siamon's presentation of Gezer to Israel

Libyan rule: the 22nd and 23rd dynasties. The fifth king of the 21st dynasty, Osorkon I (ruled *c.* 979–*c.* 973 BC), was of Libyan descent and probably was an ancestor of the 22nd dynasty, which followed a generation later. From his time to the 26th dynasty, leading Libyans in Egypt kept their Libyan names and ethnic identity, but in a spirit of ethnicity rather than cultural separatism. Although political institutions were different from those of the New Kingdom, the Libyans were culturally Egyptian, retaining only their group identity, names, and perhaps a military ethos. Toward the end of the 21st dynasty the Libyan leader of Bubastis, the great Meshwesh chief Sheshonk I (the biblical Shishak), secured special privileges from King Psusennes II (ruled *c.* 964–*c.* 950 BC) and the oracle of Amon for the mortuary cult of his father at Abydos. The oracle proffered good wishes not only for Sheshonk and his family but, significantly, also for his army. With a strong military backing, Sheshonk eventually took the throne. His reign (*c.* 950–929 BC) marks the founding of the 22nd dynasty (*c.* 950–*c.* 730 BC). Military controls were established, with garrisons under Libyan commandants serving to quell local insurrections, so that the structure of the state became more feudalistic. The dynasty tried to cement relations with Thebes through political marriages with priestly families. King Sheshonk's son Osorkon married Psusennes II's daughter, and their son eventually became high priest at Karnak. By installing their sons as high priests and promoting such marriages, kings strove to overcome the administrative division of the country. But frequent conflicts arose over the direct appointment of the Theban high priest from among the sons of Libyan kings and over the inheritance of the post by men of mixed Theban and Libyan descent. This tension took place against a background of Theban resentment of the northern dynasty. During the reign of Takelot II, strife concerning the high priestship led to civil war at Thebes. The King's son Osorkon was appointed high priest, and he achieved some semblance of order during his visits to Thebes, but he was driven from the post several times.

Renewed influence in Palestine

The initially successful 22nd dynasty revived Egyptian influence in Palestine. After Solomon's death (*c.* 936), Sheshonk I entered Palestine and plundered Jerusalem. Prestige from this exploit may have lasted through the reign of Osorkon II (formerly numbered I; ruled *c.* 929–*c.* 914 BC). In the reign of Osorkon III (ruled *c.* 888–*c.* 860 BC), Peywed Libyans posed a threat to the western Delta, perhaps necessitating a withdrawal from Palestine.

The latter part of the dynasty was marked by fragmentation of the land: Libyan great chiefs ruled numerous local areas, and there were as many as six kings in the land at a time. Increased urbanization accompanied this fragmentation, which was most intense in the Delta. Meanwhile, in Thebes, a separate 23rd dynasty was recognized.

From the 9th century BC a local Cushite state, which looked to Egyptian traditions from the colonial period of the New Kingdom, arose in the Sudan and developed around the old regional capital of Napata. The earliest ruler of the state known by name was Alara, whose piety toward Amon is mentioned in several inscriptions. His successor, Kashta, proceeded into Upper Egypt, forcing Osorkon IV (ruled *c.* 777–*c.* 750 BC) to retire to the Delta. Kashta assumed the title of king and compelled Osorkon IV's daughter Shepenwepe I, the God's Wife of Amon at Thebes, to adopt his own daughter Amonirdis I as her successor. The Cushites stressed the role of the God's Wife of Amon, who was a virgin and the consecrated partner of Amon, and sought to bypass the high priests.

The 24th and 25th dynasties. Meanwhile, the eastern Delta capital, Tanis, lost its importance to Sais in the western Delta. A Libyan prince of Sais, Tefnakhte, attempting to gain control over all Egypt, proceeded southward to Heracleopolis after acquiring Memphis. This advance was met by the Cushite ruler Piye (now the accepted reading of "Piankhi," ruled *c.* 750–*c.* 719 BC), who executed a raid as far north as Memphis and received the submission of the northern rulers (in about 730 BC). In his victory stela, Piye is portrayed as conforming strictly to Egyptian norms and reasserting traditional values against contemporary decay.

Upon Piye's return to Cush, Tefnakhte reasserted his

authority in the north, where he was eventually succeeded by his son Bocchoris, according to Manetho the sole king of the 24th dynasty (*c.* 722–*c.* 715 BC). Piye's brother Shabaka meanwhile founded the rival 25th dynasty and brought all Egypt under his rule (*c.* 719–703 BC). He had Bocchoris burned alive and removed all other claimants to the kingship.

Growth of the Assyrian Empire

In this period Egypt's internal politics were affected by the growth of the Assyrian Empire. In Palestine and Syria frequent revolts against Assyria were aided by Egyptian forces. Against the power of Assyria, the Egyptian and Nubian forces met with little success, partly because of their own fragmented politics and divided loyalties.

Although the earlier years of King Taharqa (ruled 690–664 BC), who as second son of Shabaka had succeeded his brother Shebitku (ruled 703–690 BC), were prosperous, the confrontation with Assyria became acute. In 671 BC the Assyrian king Esarhaddon entered Egypt and drove Taharqa into Upper Egypt. Two years later Taharqa regained a battered Memphis, but in 667 BC Esarhaddon's successor, Ashurbanipal, forced Taharqa to Thebes, where the Cushites held ground. Taharqa's successor, Tanutamon, defeated at Memphis a coalition of Delta princes who supported Assyria, but Ashurbanipal's reaction to this was to humiliate Thebes, which the Assyrians plundered. By 656 the Cushites had withdrawn from the Egyptian political scene, although Cushite culture survived in the Sudanese Napatan and Meroitic kingdom for another millennium.

The Late Period (664–332 BC). *The 26th dynasty (664–525 BC).* Assyria, unable to maintain a large force in Egypt, supported several Delta vassal princes, including the powerful Psamtik I of Sais. But the Assyrians faced serious problems closer to home, and Psamtik (or Psammetichus I, ruled 664–610 BC) was able to assert his independence and extend his authority as king over all Egypt without extensive use of arms, inaugurating the Saite 26th dynasty. In 656 Psamtik I compelled Thebes to submit. He allowed its most powerful man, who was Montemhat, the mayor and the fourth prophet of Amon, to retain his post and, in order to accommodate pro-Cushite sentiments, he allowed the God's Wife of Amon and the Votaress of Amon (the sister and daughter of the late king Taharqa) to remain. Psamtik I's own daughter Nitocris was adopted by the Votaress of Amon and thus became heiress to the position of God's Wife. Essential to the settling of internal conflicts was the Saite dynasty's superior army, composed of Libyan soldiers, whom the Greeks called Machimoi (warriors), and Greek and Carian mercenaries, who formed part of the great emigration from the Aegean in the 7th and 6th centuries BC. Greek pirates raiding the Delta coast were induced by Psamtik I to serve in his army and were settled like the Machimoi in colonies at the Delta's strategically important northeastern border. Trade developed between Egypt and Greece, and more Greeks settled in Egypt.

Foreign policy under the Saites

The Saite dynasty generally pursued a foreign policy that avoided territorial expansion and tried to preserve the status quo. Assyria's power was waning. In 655 BC Psamtik I marched into Philistia in pursuit of the Assyrians, and in 620 BC he apparently repulsed Scythians from the Egyptian frontier. During the reign of his son Necho II (610–595 BC), Egypt supported Assyria as a buffer against the potential threat of the Medes and the Babylonians. Necho was successful in Palestine and Syria until 605 BC, when the Babylonian Nebuchadrezzar inflicted a severe defeat on Egyptian forces at Carchemish. After withdrawing his troops from Asia, Necho concentrated on developing Egyptian commerce; the grain that was delivered to Greece was paid for in silver. He also built up the navy and began a canal linking the Nile with the Red Sea. Under Psamtik II (ruled 595–589 BC) there was a campaign through the Napatan kingdom involving the use of Greek and Carian mercenaries who left their inscriptions at Abu Simbel; at the same time the names of the long-dead Cushite rulers were erased from their monuments in Egypt. Psamtik II also made an expedition to Phoenicia accompanied by priests; whether it was a military or a goodwill mission is unknown.

The next king, Apries (ruled 589–570 BC), tried unsuccessfully to end Babylonian domination of Palestine and Syria. With the withdrawal of Egyptian forces, Nebuchadrezzar destroyed the temple in Jerusalem in 586 BC. In the aftermath of his conquest, many Jews fled to Egypt, where some were enlisted as soldiers in the Persian army of occupation. Apries' army was then defeated in Libya when it attacked the Greek colony at Cyrene, some 620 miles west of the Delta; this led to an army mutiny and to civil war in the Delta. A new Saite king, Amasis (or Ahmose II; ruled 570–526 BC), usurped the throne and drove Apries into exile. Two years later Apries invaded Egypt with Babylonian support, but he was defeated and killed by Amasis, who nonetheless buried him with full honours. Amasis returned to a more conservative foreign policy in a long, prosperous reign. To reduce friction between Greeks and Egyptians, especially in the army, Amasis withdrew the Greeks from the military colonies and transferred them to Memphis, where they formed a sort of royal bodyguard. He limited Greek trade in Egypt to Sais, Memphis, and Naukratis, the latter becoming the only port to which Greek wares could be brought, so that taxes on imports and on business could be enforced. Naukratis prospered and Amasis was seen by the Greeks as a benefactor. In foreign policy he supported a waning Babylonia, now threatened by Persia; but six months after his death in 526 BC the Persian Cambyses II (ruled as pharaoh 525–522 BC) penetrated Egypt, reaching Nubia in 525.

Nationalism and cultural revival

As was common in the Near East in this period, the Saite kings used foreigners as mercenaries to prevent foreign invasions. An element within Egyptian culture, however, resisted any influence of the resident foreigners and gave rise to a nationalism that provided psychological security in days of political uncertainty. A cultural revival was initiated in the 25th dynasty and continued throughout the 26th. Temples and the priesthood were overtly dominant. In their inscriptions the elite displayed their priestly titles but did not mention the administrative roles that they probably also performed. Throughout the country, people of substance dedicated land to temple endowments that supplemented royal donations. The god Seth, who had been an antithetic element in Egyptian religion, came gradually to be proscribed as the god of foreign lands.

The revival of this period was both economic and cultural, but there is less archaeological evidence preserved than for earlier times because the economic centre of the country was now the Delta, where conditions for the preservation of ancient sites were unfavourable. Prosperity increased throughout the 26th dynasty, reaching a high point in the reign of Amasis. Temples throughout the land were added to, often in hard stones carved with great skill. The chief memorials of private individuals were often temple statues, of which many fine examples were dedicated, again mostly in hard stones. In temple and tomb decoration and in statuary, the Late Period rejected its immediate predecessors and looked to the great periods of the past for models. There was, however, also significant innovation. In writing, the demotic script, the new cursive form, was introduced from the north and spread gradually through the country. Demotic wrote a contemporary form of the language, and administrative Late Egyptian disappeared. Hieratic was, however, retained for literary and religious texts, among which very ancient material, such as the Pyramid Texts, was revived and inscribed in tombs and on coffins and sarcophagi.

The Late Period saw the greatest development of animal worship in Egypt. This feature of religion, which was the subject of much interest and scorn among classical writers, had always existed but had been of minor importance. In the Late and Ptolemaic periods, it became one of the principal forms of popular religion in an intensely religious society. Many species of animal were mummified and buried, and towns sprang up in the necropolises to cater for the needs of dead animals and their worshipers. At Ṣaqqārah the Apis bull, which had been worshiped as a manifestation of the god Ptah since the 1st dynasty, was buried in a huge granite sarcophagus in ceremonies in which royalty might take part. At least 10 species, from ibises, buried by the million, to dogs, were buried by the

heterogeneous population of Memphis, Egypt's largest city.

Egypt under Achaemenid rule. *The 27th dynasty.* According to the Greek historian Herodotus, who visited Egypt around 450 BC, Cambyses II's conquest of Egypt was ruthless and sacrilegious. Contemporary Egyptian sources, however, treat him in a more favourable light. He assumed the full titulary of an Egyptian king and paid honour to the goddess Neith of Sais. His unfavourable later reputation probably resulted from adverse propaganda by Egyptian priests, who resented his reduction of temple income. Darius I, who succeeded Cambyses in 522 BC and ruled as pharaoh until 486 BC, was held in higher esteem because he was concerned with improving the temples and restored part of their income, and because he codified laws as they had been in the time of Amasis. These stances, which aimed to win over priests and learned Egyptians, were elements of his strategy to retain Egypt as a lasting part of the Persian Empire. Egypt, together with the Libyan oases and Cyrenaica, formed the sixth Persian satrapy (province), whose satrap resided at Memphis, while Persian governors under him held posts in cities throughout the land. Under Darius I the tax burden upon Egyptians was relatively light, and Persians aided Egypt's economy through irrigation projects and improved commerce, enhanced by the completion of the canal to the Red Sea.

The Persian defeat by the Athenians at Marathon in 490 BC had significant repercussions in Egypt. On Darius I's death in 486 BC a revolt broke out in the Delta, perhaps instigated by Libyans of the west Delta. The result was that the Persian king Xerxes reduced Egypt to the status of a conquered province. Egyptians dubbed him the "criminal Xerxes." He never visited Egypt and appears not to have utilized Egyptians in high positions in the administration. Xerxes' murder in 465 BC was the signal for another revolt in the western Delta. It was led by a dynast, Inaros, who acquired control over the Delta and was supported by Athenian forces against the Persians. Inaros was crucified by the Persians in 454 BC, when they regained control of most of the Delta. In the later 5th century BC, under the rule of Artaxerxes I (ruled as pharaoh 465–424 BC) and Darius II (ruled as pharaoh 424–404 BC), conditions in Egypt were very unsettled, and scarcely any monuments of the period have been identified.

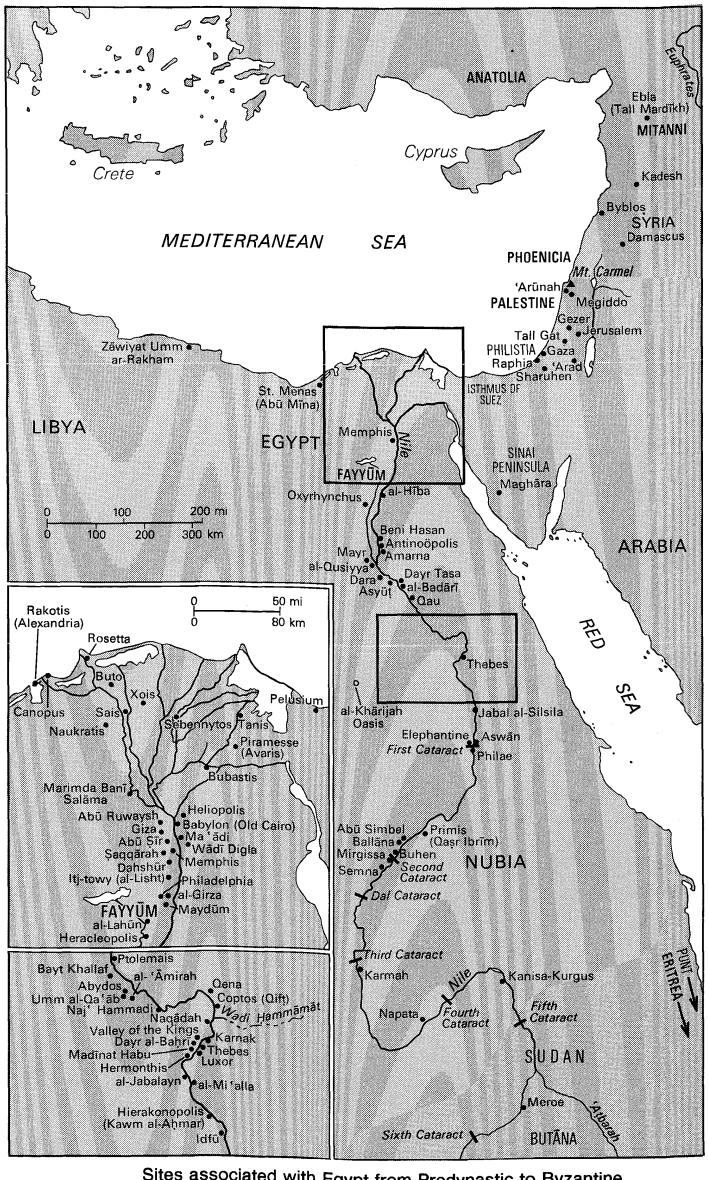
The 28th, 29th, and 30th dynasties. The death of Darius II in 404 BC prompted a successful rebellion in the Delta, and the Egyptian Amyrtaeus formed a Saite 28th dynasty, of which he was the sole king (404–399 BC). His rule was recognized in Upper Egypt by 401 BC, at a time when Persia's troubles elsewhere forestalled an attempt to regain Egypt.

Despite growing prosperity and success in retaining independence, 4th-century Egypt was characterized by continual internal struggle for the throne. After a long period of fighting in the Delta, a 29th dynasty (399–380 BC) emerged from Mendes. Achoris (ruled 393–380 BC), its third and final ruler, was especially vigorous, and the prosperity of his reign is indicated by many monuments in Upper and Lower Egypt. Once again Egypt was active in international politics, forming alliances with the opponents of Persia and building up its army and navy. The Egyptian army included Greeks both as mercenaries and as commanders; the mercenaries were not permanent residents of military camps in Egypt but native Greeks seeking payment for their services in gold. Payment was normally made in non-Egyptian coins, because as yet Egypt had no coinage in general circulation; the foreign coins may have been acquired in exchange for exports of grain, papyrus, and linen. Some Egyptian coins were minted in the 4th century, but they do not seem to have gained widespread acceptance.

Aided by the Greek commander Chabrias of Athens and his elite troops, Achoris prevented a Persian invasion; but after Achoris' death in 380 BC his son Nepherites II lasted only four months before a general, Nectanebo I (Nekhtnebef; ruled 380–362 BC) of Sebennytos, usurped the throne, founding the 30th dynasty (380–343 BC). In 373 BC the Persians attacked Egypt, and, although Egyptian losses were heavy, disagreement between the Persian satrap Pharnabazus and his Greek commander over strategy, combined with a timely inundation of the Delta, saved the day for Egypt. With the latent dissolution of the Persian Empire under the weak Artaxerxes II, Egypt was relatively safe from further invasion; it remained prosperous throughout the dynasty.

Egypt had a more aggressive foreign policy under Nectanebo's son Tachos (ruled *c.* 365–360 BC). Possessing a strong army and navy composed of Egyptian Machimoi and Greek mercenaries and supported by Chabrias and the Spartan king Agesilaus, Tachos (in Egyptian called Djeho) invaded Palestine. But friction between Tachos and Agesilaus and the cost of financing the venture proved to be Tachos' undoing. In an attempt to raise funds quickly, he had imposed taxes and seized temple property. Egyptians, especially the priests, resented this burden and supported Tachos' nephew Nectanebo II (Nekhtharehbe; ruled 360–343 BC) in his usurpation of the throne. The cost of retaining the allegiance of mercenaries proved too high for a nonmonetary economy.

Agesilaus supported Nectanebo in his defensive foreign policy, and the priests sanctioned the new king's building activities. Meanwhile, Persia enjoyed a resurgence under Artaxerxes III (Ochus); but a Persian attack upon Egypt in 350 BC was repulsed. In 343 BC the Persians once again marched against Egypt. The first battle was fought at Pelusium and proved the superiority of Persia's strategy.



Sites associated with Egypt from Predynastic to Byzantine times.

Eventually the whole Delta, then the rest of Egypt, fell to Artaxerxes III, and Nectanebo fled to Nubia.

The 4th century BC was the last flourishing period of an independent Egypt and saw notable artistic and literary achievements. The 26th dynasty artistic revival evolved further toward more complex forms that culminated briefly in a Greco-Egyptian stylistic fusion, as seen in the tomb of Petosiris at Tūnah al-Jabal from the turn of the 3rd century BC. In literature works continued to be transmitted, and possibly composed, in hieratic, but that tradition was to develop no further. Demotic literary works began to appear, including stories set in the distant past, mythological tales, and an acrostic text apparently designed to teach an order of sounds in the Egyptian language.

Return of Persian rule

The second Persian period. Artaxerxes dealt harshly with Egypt, razing city walls, rifling temple treasuries, and removing sacred books. Persia acquired rich booty in its determination to prevent Egypt from further rebelling. After the murder of Artaxerxes III, in 338 BC, there was a brief obscure period during which a Nubian prince, Khabbash, seems to have gained control over Egypt, but Persian domination was reestablished in 335 BC under Darius III Codommanus. It was to last only three years.

(E.F.W./J.R.Ba.)

MACEDONIAN AND PTOLEMAIC EGYPT (332–30 BC)

The Macedonian conquest. In the autumn of 332 BC Alexander the Great invaded Egypt with his mixed army of Macedonians and Greeks and found the Egyptians ready to throw off the oppressive control of the hated Persians.

Alexander's welcome in Egypt

Alexander was welcomed by the Egyptians as a liberator and took the country without a battle. He journeyed to Siwa Oasis in the Western Desert to visit the Oracle of Amon, renowned in the Greek world; it disclosed the information that Alexander was the son of Amon. There may also have been a coronation at the Egyptian capital, Memphis, which, if it occurred, would have placed him firmly in the tradition of the pharaohs; the same purpose may be seen in the later dissemination of the romantic myth that gave him an Egyptian parentage by linking his mother, Olympias, with the last pharaoh, Nectanebo II.

Alexander left Egypt in the spring of 331 BC, dividing the military command between Balacrus, son of Amyntas, and Peucestas, son of Makartatos. The earliest known Greek documentary papyrus, found at Ṣaqqārah in 1973, reveals the sensitivity of the latter to Egyptian religious institutions in a notice that reads: "Order of Peucestas. No-one is to pass. The chamber is that of a priest." The civil administration was headed by an official with the Persian title of satrap, one Cleomenes of Naukratis. When Alexander died in 323 BC and his generals divided his empire, the position of satrap was claimed by Ptolemy, son of a Macedonian nobleman named Lagus. The senior general Perdiccas, the holder of Alexander's royal seal and prospective regent for Alexander's posthumous son, might well have regretted his failure to take Egypt. He gathered an army and marched from Asia Minor to wrest Egypt from Ptolemy in 321 BC; but Ptolemy had Alexander's corpse, Perdiccas' army was not wholehearted in support, and the Nile crocodiles made a good meal from the flesh of the invaders.

The Ptolemaic dynasty. Until the day when he openly assumed an independent kingship as Ptolemy I Soter, on Nov. 7, 305 BC, Ptolemy used only the title satrap of Egypt, but the great hieroglyphic Satrap stela, which he had inscribed in 311 BC, indicates a degree of self-confidence that transcends his viceregal role. It reads, "I, Ptolemy the satrap, I restore to Horus, the avenger of his father, the lord of Pe and to Buto, the lady of Pe and Dep, the territory of Patanut, from this day forth for ever, with all its villages, all its towns, all its inhabitants, all its fields." The inscription emphasizes Ptolemy's own role in wresting the land from the Persians (though the epithet of Soter, meaning "Saviour," resulted not from his actions in Egypt but from the gratitude of the people of Rhodes for his having relieved them from a siege in 315 BC) and links him with Khabbash, who had laid claim to the kingship during the last Persian occupation in about 338 BC.

Egypt was ruled by Ptolemy's descendants until the death of Cleopatra VII on Aug. 12, 30 BC. The kingdom was one of several that emerged in the aftermath of Alexander's death and struggles of his successors. It was the wealthiest, however, and, for much of the next 300 years, the most powerful politically and culturally, and it was the last to fall directly under Roman dominion. In many respects, the character of the Ptolemaic monarchy in Egypt set a style for other Hellenistic kingdoms; this style emerged from the Greeks' and Macedonians' awareness of the need to dominate Egypt, its resources, and its people and at the same time to turn the power of Egypt firmly toward the context of a Mediterranean world that was becoming steadily more Hellenized.

The Ptolemies (305–145 BC). The first 160 years of the Ptolemaic dynasty are conventionally seen as its most prosperous era. Little is known of the foundations laid in the reign of Ptolemy I Soter (304–282 BC), but the increasing amount of documentary, inscriptional, and archaeological evidence from the reign of his son and successor, Ptolemy II Philadelphus (285–246 BC), shows that the kingdom's administration and economy underwent a thorough reorganization. A remarkable demotic text of the year 258 BC refers to orders for a complete census of the kingdom that was to record the sources of water; the position, quality, and irrigation potential of the land; the state of cultivation; the crops grown; and the extent of priestly and royal landholdings. There were important agricultural innovations in this period. New crops were introduced, and massive irrigation works brought under cultivation a great deal of new land, especially in the Fayyūm, where many of the immigrant Greeks were settled.

Macedonian-Greek character of the monarchy

The Macedonian-Greek character of the monarchy was vigorously preserved. There is no more emphatic sign of this than the growth and importance of the city of Alexandria. It had been founded, on a date traditionally given as April 7, 331 BC, by Alexander the Great on the site of the insignificant Egyptian village of Rakotis in the northwestern Delta, and it ranked as the most important city in the eastern Mediterranean until the foundation of Constantinople in the 4th century AD. The importance of the new Greek city was soon emphasized by contrast to its Egyptian surroundings when the royal capital was transferred, within a few years of Alexander's death, from Memphis to Alexandria. The Ptolemaic court cultivated extravagant luxury in the Greek style in its magnificent and steadily expanding palace complex, which occupied as much as a third of the city by the early Roman period. Its grandeur was emphasized in the reign of Ptolemy II Philadelphus by the foundation of a quadrennial festival, the Ptolemaieia, which was intended to enjoy a status equal to that of the Olympic Games. The festival was marked by a procession of amazingly elaborate and ingeniously constructed floats, with scenarios illustrating Greek religious cults.

Ptolemy II gave the dynasty another distinctive feature when he married his full sister, Arsinoe II, one of the most powerful and remarkable women of the Hellenistic age. They became, in effect, co-rulers, and both took the epithet Philadelphus ("Brother-Loving" and "Sister-Loving"). The practice of consanguineous marriage was followed by most of their successors and imitated by ordinary Egyptians too, even though it had not been a standard practice in the pharaonic royal houses and had been unknown in the rest of the native Egyptian population. Arsinoe played a prominent role in the formation of royal policy. She was displayed on the coinage and was eventually worshiped, perhaps even before her death, in the distinctively Greek style of ruler cult that developed in this reign.

Ptolemaic empire

From the first phase of the wars of Alexander's successors the Ptolemies had harboured imperial ambitions. Ptolemy I won control of Cyprus and Cyrene and quarreled with his neighbour over control of Palestine. In the course of the 3rd century a powerful Ptolemaic empire developed, which, for much of the period, laid claim to sovereignty in the Levant, in many of the cities of the western and southern coast of Asia Minor, in some of the Aegean islands, and in a handful of towns in Thrace, as well as in Cyprus and Cyrene. Family connections and dynastic alliances, especially between the Ptolemies and the neigh-

bouring Seleucids, played a very important role in these imperialistic ambitions. Such links were far from able to preserve harmony between the royal houses (between 274 and 200 BC five wars were fought with the Seleucids over possession of territory in Syria and the Levant), but they did keep the ruling houses relatively compact, interconnected, and more true to their Macedonian-Greek origins.

When Ptolemy II Philadelphus died in 246 BC, he left a prosperous kingdom to his successor, Ptolemy III Euergetes (246–222 BC). His reign saw a very successful campaign against the Seleucids in Syria, occasioned by the murder of Euergetes' sister, Berenice, who had been married to the Seleucid Antiochus II. To avenge Berenice, Euergetes marched into Syria, where he won a great victory. He gained popularity at home by recapturing statues of Egyptian gods originally taken by the Persians. The decree promulgated at Canopus in the Delta on March 4, 238 BC, attests both this event and the many great benefactions conferred on Egyptian temples throughout the land. It was during Euergetes' reign, for instance, that the rebuilding of the great Temple of Horus at Idfū (Apollinopolis Magna) was begun.

Euergetes was succeeded by his son Ptolemy IV Philopator (222–205 BC), whom the Greek historians portray as a weak and corrupt ruler, dominated by a powerful circle of Alexandrian Greek courtiers. The reign was notable for another serious conflict with the Seleucids, which ended in 217 BC in a great Ptolemaic victory at Raphia in southern Palestine. The battle is notable for the fact that large numbers of native Egyptian soldiers fought alongside the Macedonian and Greek contingents. Events surrounding the death of Philopator and the succession of the youthful Ptolemy V Epiphanes (205–180 BC) are obscured by court intrigue. Before Epiphanes had completed his first decade of rule, serious difficulties arose. Native revolts in the south, which had been sporadic in the second half of the 3rd century, became serious and weakened the hold of the monarch on a vital part of the kingdom. These revolts, which produced native claimants to the kingship, are generally attributed to the native Egyptians' realization, after their contribution to the victory at Raphia, of their potential power. Trouble continued to break out for several more decades. By about 196 a great portion of the Ptolemaic overseas empire had been permanently lost (though there may have been a brief revival in the Aegean islands in about 165–145 BC). To shore up and advertise the strength of the ruling house at home and abroad, the administration adopted a series of grandiloquent honorific titles for its officers. To conciliate Egyptian feelings, a religious synod that met in 196 to crown Epiphanes at Memphis (the first occasion on which a Ptolemy is certainly known to have been crowned at the traditional capital) decreed extensive privileges for the Egyptian temples, as recorded on the Rosetta Stone.

Loss of the overseas empire

The reign of Ptolemy VI Philometor (180–145 BC), a man of pious and magnanimous character, was marked by renewed conflict with the Seleucids after the death of his mother, Cleopatra I, in 176 BC. In 170 BC Antiochus IV of Syria invaded Egypt and established a protectorate; in 168 BC he returned, accepted coronation at Memphis, and installed a Seleucid governor. But he had failed to reckon with more powerful interests: those of Rome. In the summer of 168 BC a Roman ambassador, Popillius Laenas, arrived at Antiochus' headquarters near Pelusium in the Delta and staged an awesome display of Roman power. He ordered Antiochus to withdraw from Egypt. Antiochus asked for time to consult his advisers. Laenas drew a circle around the King with his stick and told him to answer before he stepped out of the circle. Only one answer was possible, and by the end of July Antiochus had left Egypt. Philometor's reign was further troubled by rivalry with his brother, later Ptolemy VIII Euergetes II Physcon. The solution, devised under Roman advice, was to remove Physcon to Cyrene, where he remained until Philometor died in 145 BC; but it is noteworthy that in 155 BC Physcon took the step of bequeathing the kingdom of Cyrene to the Romans in the event of his untimely death.

Dynastic strife and decline (145–30 BC). Physcon was able to rule in Egypt until 116 BC with his sister Cleopatra II (except for a period in 131–130 BC when she was in revolt) and her daughter Cleopatra III. His reign was marked by generous benefactions to the Egyptian temples, but he was detested as a tyrant by the Greeks, and the historical accounts of the reign emphasize his stormy relations with the Alexandrian populace.

During the last century of Ptolemaic rule, Egypt's independence was exercised under Rome's protection and at Rome's discretion. For much of the period Rome was content to support a dynasty that had no overseas possession except Cyprus after 96 BC (the year in which Cyrene was bequeathed to Rome by Ptolemy Apion) and no ambitions threatening Roman interests or security. After a series of brief and unstable reigns, Ptolemy XII Auletes acceded to the throne in 80 BC. He maintained his hold for 30 years, despite the attractions that Egypt's legendary wealth held for avaricious Roman politicians. In fact, Auletes had to flee Egypt in 58 BC and was restored by Pompey's friend Gabinius in 55 BC, no doubt after spending so much in bribes that he had to bring back Rabirius Postumus, one of his Roman creditors, to Egypt with him to manage his financial affairs.

In 52 BC, the year before his death, Auletes associated with himself on the throne his daughter Cleopatra VII and his elder son Ptolemy XIII (who died in 47 BC). The reign of Cleopatra was that of a vigorous and exceptionally able queen who was ambitious, among other things, to revive the prestige of the dynasty by cultivating influence with powerful Roman commanders and using their capacity to aggrandize Roman clients and allies. Julius Caesar pursued Pompey to Egypt in 48 BC. After learning of Pompey's murder at the hands of Egyptian courtiers, Caesar stayed long enough to enjoy a sightseeing tour up the Nile in the Queen's company in the summer of 47 BC. When he left for Rome, Cleopatra was pregnant with a child she claimed was Caesar's. The child, a son, was named Caesarion ("Little Caesar"). Cleopatra and Caesarion later followed Caesar back to Rome but, after his assassination in 44 BC, they returned hurriedly to Egypt and she tried for a while to play a neutral role in the struggles between the Roman generals and their factions.

Reign of Cleopatra VII

Her long liaison with Mark Antony began when she visited him at Tarsus in 41 BC and he returned to Egypt with her. Between 36 and 30 BC the famous romance between the Roman general and the eastern queen was exploited to great effect by Antony's political rival Octavian. By 34 BC Caesarion was officially co-ruler with Cleopatra, but his rule clearly was an attempt to exploit the popularity of Caesar's memory. In the autumn Cleopatra and Antony staged an extravagant display in which they made grandiose dispositions of territory in the east to their children, Alexander Helios, Ptolemy, and Cleopatra Selene. Cleopatra and Antony were portrayed to the Roman public as posing for artists in the guise of Dionysus and Isis or whiling away their evenings in rowdy and decadent banquets that kept the citizens of Alexandria awake all night. But this propaganda war was merely the prelude to armed conflict, and the issue was decided in September 31 BC in a naval battle at Actium in western Greece. When the battle was at its height Cleopatra and her squadron withdrew, and Antony eventually followed suit. They fled to Alexandria but could do little more than await the arrival of the victorious Octavian 10 months later. Alexandria was captured and Antony and Cleopatra committed suicide—he by falling on his sword, she probably by the bite of an asp—in August of 30 BC. It is reported that when Octavian reached the city he visited and touched the preserved corpse of Alexander the Great, causing a piece of the nose to fall off. He refused to gaze upon the remains of the Ptolemies, saying "I wished to see a king, not corpses."

Government and conditions under the Ptolemies. The changes brought to Egypt by the Ptolemies were momentous; the land's resources were harnessed with unparalleled efficiency and the result was that it became the wealthiest of the Hellenistic kingdoms. Land under cultivation was increased, new crops were introduced (especially important was the introduction of naked tetraploid wheat, *triticum durum,* to replace the traditional husked emmer,

Improvements to agriculture

triticum dicoccum). The population, estimated at perhaps 3,000,000–4,000,000 in the Late Dynastic Period, may have more than doubled by the early Roman period to a figure of 7,500,000 or 8,000,000, a level not reached again until the late 19th century. Some of the increase was due to immigration; particularly during the 2nd and 3rd centuries many settlers were attracted from the cities of Asia Minor and the Greek islands, as well as large numbers of Jews from Palestine. The flow may have decreased later in the Ptolemaic period, and it is often suggested, on slender evidence, that there was a serious decline in prosperity in the 1st century BC. If so, there may have been some reversal of this trend under Cleopatra VII.

Administration. The foundation of the prosperity was the governmental system devised to exploit the country's economic resources. Directly below the monarch were a handful of powerful officials whose competence extended over the entire land: a chief finance minister, a chief accountant, and a chancery of ministers in charge of records, letters, and decrees. A level below them lay the broadening base of a pyramid of subordinate officials with competence in limited areas, which extended down to the chief administrator of each individual village (*kōmarchēs*). Between the chief ministers and the village officials stood those such as the nome-steward (*oikonomos*) and *stratēgoi,* whose competence extended over one of the more than 30 nomes of Egypt, the long-established geographic divisions. In theory this bureaucracy could regulate and control the economic activities of every subject in the land, its smooth operation guaranteed by the multiplicity of officials capable of checking each upon the other. In practice, it is difficult to see a rigid civil-service mentality at work, involving clear demarcation of departments; specific functions might well have been performed by different officials according to local need and the availability of a person competent to take appropriate action.

By the same token, rigid lines of separation between military and civil, legal and administrative matters are difficult to perceive. The same official might perform duties in one or all of these areas, and the law in particular regulated every activity to an extent that the use of the terms legal and judicial tends to hide. The military was inevitably integrated into civilian life because its soldiers were also farmers who enjoyed royal grants of land, either as Greek cleruchs (holders of allotments) with higher status and generous grants, or as native Egypt *machimoi* with small plots. Interlocking judiciary institutions, in the form of Greek and Egyptian courts (*chrēmatistai* and *laokritai*), provided the means for Greeks and Egyptians to regulate their legal relationships according to the language in which they conducted their business. The bureaucratic power was heavily weighted in favour of the Greek speakers, the dominant elite. Egyptians were nevertheless able to obtain official posts in the bureaucracy, gradually infiltrating to the highest levels, but in order to do so they had to Hellenize.

Economy. The basis of Egypt's legendary wealth was the highly productive land, which technically remained in royal ownership. A considerable portion was kept under the control of temples, and the remainder was leased out on a theoretically revocable basis to tenant-farmers. A portion also was available to be granted as gifts to leading courtiers; one of these was Apollonius, the finance minister of Ptolemy II Philadelphus, who had an estate of 10,000 arourae (about 6,500 acres) at Philadelphia in the Fayyūm. Tenants and beneficiaries were able to behave very much as if these leases and grants were private property. The revenues in cash and kind were enormous, and royal control extended to the manufacture and marketing of almost all important products, including papyrus, oil, linen, and beer. An extraordinarily detailed set of revenue laws, promulgated under Ptolemy II Philadelphus, laid down rules for the way in which officials were to monitor the production of such commodities. In fact, the Ptolemaic economy was very much a mixture of direct royal ownership and exploitation by private enterprise under regulated conditions.

Control of land

One fundamental and far-reaching Ptolemaic innovation was the systematic monetarization of the economy. This too the monarchy controlled from top to bottom by operating a closed monetary system, which permitted only the royal coinage to circulate within Egypt. A sophisticated banking system underpinned this practice, operating again with a mixture of direct royal control and private enterprise and handling both private financial transactions and those that directed money into and out of the royal coffers. One important concomitant of this change was an enormous increase in the volume of trade, both within Egypt and abroad, which eventually reached its climax under the peaceful conditions of Roman rule. Here the position and role of Alexandria as the major port and trading entrepôt was crucial: the city handled a great volume of Egypt's domestic produce, as well as the import and export of luxury goods to and from the East and the cities of the eastern Mediterranean. It developed its own importance as an artistic centre, the products of which found ready markets throughout the Mediterranean. Alexandrian glassware and jewelry were particularly fine; Greek-style sculpture of the late Ptolemaic period shows especial excellence; and it is likely that the city was also the major production centre for high-quality mosaic work.

Alexandria's importance to trade

Religion. The Ptolemies were powerful supporters of the native Egyptian religious foundations, the economic and political power of which was, however, carefully controlled. A great deal of the building and restoration work in many of the most important Egyptian temples is Ptolemaic, particularly from the period of about 150–50 BC, and the monarchs appear on temple reliefs in the traditional forms of the Egyptian kings. The native traditions persisted in village temples and local cults, many having particular associations with species of sacred animals or birds. At the same time, the Greeks created their own identifications of Egyptian deities, identifying Amon with Zeus, Horus with Apollo, Ptah with Hephaestus, and so on. They also gave some deities, such as Isis, a more universal significance that ultimately resulted in the spread of her mystery cult throughout the Mediterranean world. The impact of the Greeks is most obvious in two phenomena. One is the formalized royal cult of Alexander and the Ptolemies, which evidently served both a political and a religious purpose. The other is the creation of the cult of Sarapis, which at first was confined to Alexandria but soon became universal. The god was represented as a Hellenized deity and the form of cult is Greek; but its essence is the old Egyptian notion that the sacred Apis bull merged its divinity in some way with the god Osiris when it died.

Greek identifications of Egyptian deities

Culture. The continuing vitality of the native Egyptian artistic tradition is clearly and abundantly expressed in the temple architecture and the sculpture of the Ptolemaic period. The Egyptian language continued in use in its hieroglyphic and demotic forms until late in the Roman period, and it survived through the Byzantine period and beyond in the form of Coptic. The Egyptian literary tradition flourished vigorously in the Ptolemaic period and produced a large number of works in demotic. The genre most commonly represented is the romantic tale, exemplified by several story cycles, which are typically set in the native, Pharaonic milieu and involve the gods, royal figures, magic, romance, and the trials and combats of heroes. Another important category is the Instruction Text, the best known of the period being that of Ankhsheshonq, which consists of a list of moralizing maxims, composed, as the story goes, when Ankhsheshonq was imprisoned for having failed to inform the pharaoh of an assassination plot. Another example, known as Papyrus Insinger, is a more narrowly moralizing text. But the arrival of a Greek-speaking elite had an enormous impact on cultural patterns. The Egyptian story cycles were probably affected by Greek influence; literary and technical works were translated into Greek; and under royal patronage an Egyptian priest named Manetho of Sebennytos wrote an account of the kings of Egypt, in Greek. Most striking is the diffusion of the works of the poets and playwrights of classical Greece among the literate Greeks in the towns and villages of the Nile Valley.

Influence of Greek on the literature

Thus there are clear signs of the existence of two interacting but distinct cultural traditions in Ptolemaic Egypt.

This was certainly reflected in a broader social context. The written sources offer little direct evidence of racial discrimination by Greeks against Egyptians, but Greek and Egyptian consciousness of the Greeks' social and economic superiority comes through strongly from time to time; intermarriage was one means, though not the only one, by which Egyptians could better their status and Hellenize. Many native Egyptians learned to speak Greek, some to write it as well; some even went so far as to adopt Greek names in an attempt to assimilate themselves to the elite group.

Hellenization of the Egyptians

Alexandria occupied a unique place in the history of literature, ideas, scholarship, and science for almost a millennium after the death of its founder. Under the royal patronage of the Ptolemies, and in an environment almost oblivious to its Egyptian surroundings, Greek culture was preserved and developed. Early in the Ptolemaic period, probably in the reign of Ptolemy I Soter, the Museum ("Shrine of the Muses") was established within the palace complex. Strabo, who saw it early in the Roman period, described it as having a covered walk, an arcade with recesses and seats, and a large house containing the dining hall of the members of the Museum, who lived a communal existence. The Great Library of Alexandria (together with its offshoot in the Sarapeum) was indispensable to the functioning of the scholarly community in the Museum. Books were collected voraciously under the Ptolemies, and at its height the library's collection probably numbered close to 500,000 papyrus rolls, most of them containing more than one work.

Poets and scholars at Alexandria

The major poets of the Hellenistic period, Theocritus, Callimachus, and Apollonius of Rhodes, all took up residence and wrote there. Scholarship flourished, preserving and ordering the manuscript traditions of much of the classical literature from Homer onward. Librarian-scholars such as Aristophanes of Byzantium and his pupil Aristarchus made critical editions and wrote commentaries and works on grammar. Also notable was the cultural influence of Alexandria's Jewish community, which is inferred from the fact that the Pentateuch was first translated into Greek at Alexandria during the Ptolemaic period. One by-product of this kind of activity was that Alexandria became the centre of the book trade, and the works of the classical authors were copied there and diffused among a literate Greek readership scattered in the towns and villages of the Nile Valley.

The Alexandrian achievement in scientific fields was also enormous. Great advances were made in pure mathematics, mechanics, physics, geography, and medicine. Euclid worked in Alexandria in about 300 BC and achieved the systematization of the whole existing corpus of mathematical knowledge and the development of the method of proof by deduction from axioms. Archimedes was there in the 3rd century BC and is said to have invented the Archimedean screw when he was in Egypt; Eratosthenes calculated the Earth's circumference and was the first to attempt a map of the world based on a system of lines of latitude and longitude; and the school of medicine founded in the Ptolemaic period retained its leading reputation into the Byzantine era. Late in the Ptolemaic period Alexandria began to develop as a great centre of Greek philosophical studies as well. In fact, there was no field of literary, intellectual, or scientific activity to which Ptolemaic Alexandria failed to make an important contribution. (A.E.S./A.K.B.)

ROMAN AND BYZANTINE EGYPT (30 BC–AD 642)

Egypt as a province of Rome. "I added Egypt to the Empire of the Roman people." With these words the emperor Augustus (as Octavian was known from 27 BC) summarized the subjection of Cleopatra's kingdom in the great inscription that records his achievements. The province was to be governed by a viceroy, a prefect with the status of a Roman knight (eques) who was directly responsible to the emperor. The first viceroy was the Roman poet and soldier Cornelius Gallus, who boasted too vaingloriously of his military achievements in the province and paid for it first with his position and then with his life. Roman senators were not allowed to enter Egypt without the emperor's permission, because this wealthiest of provinces could be held militarily by a very small force; and the threat implicit in an embargo on the export of grain supplies, vital to the provisioning of the city of Rome and its populace, was obvious. Internal security was guaranteed by the presence of three Roman legions (later reduced to two), each about 6,000 strong, and several cohorts of auxiliaries. In the first decade of Roman rule the spirit of Augustan imperialism looked farther afield, attempting expansion to the east and to the south. An expedition to Arabia by the prefect Aelius Gallus in about 26–25 BC was undermined by the treachery of the Nabataean Syllaeus, who led the Roman fleet astray in uncharted waters. Arabia was to remain an independent though friendly client of Rome until AD 106, when the emperor Trajan (ruled AD 98–117) annexed it, making it possible to reopen Ptolemy II's canal from the Nile to the head of the Gulf of Suez. To the south the Meroitic people beyond the First Cataract had taken advantage of Gallus' preoccupation with Arabia and mounted an attack on the Thebaid. The next Roman prefect, Petronius, led two expeditions into the Meroitic kingdom (*c.* 24–22 BC), captured several towns, forced the submission of the formidable queen, who was characterized by Roman writers as "the one-eyed Queen Candace," and left a Roman garrison at Primis (Qaṣr Ibrīm). But thoughts of maintaining a permanent presence in Lower Nubia were soon abandoned, and within a year or two the limits of Roman occupation had been set at Hiera Sykaminos, some 50 miles south of the First Cataract. The mixed character of the region is indicated, however, by the continuing popularity of the goddess Isis among the people of Meroe and by the Roman emperor Augustus' foundation of a temple at Kalabsha dedicated to the local god Mandulis.

Attempts to expand Rome's territory

Egypt achieved its greatest prosperity under the shadow of the Roman peace which, in effect, depoliticized it. Roman emperors or members of their families visited Egypt—Tiberius' nephew and adopted son, Germanicus; Vespasian and his elder son, Titus; Hadrian; Septimius Severus; Diocletian—to see the famous sights, receive the acclamations of the Alexandrian populace, attempt to ensure the loyalty of the volatile subjects, or initiate administrative reform. Occasionally its potential as a power base was realized. Vespasian, the most successful of the imperial aspirants in the "Year of the Four Emperors," was first proclaimed at Alexandria on July 1, AD 69, in a maneuver contrived by the prefect of Egypt, Tiberius Julius Alexander. Others were less successful. Avidius Cassius, the son of a former prefect of Egypt, revolted against Marcus Aurelius in AD 175, stimulated by false rumours of Marcus' death, but his attempted usurpation lasted only three months. For several months in AD 297/298 Egypt was under the dominion of a mysterious usurper named Lucius Domitius Domitianus. The emperor Diocletian was present at the final capitulation of Alexandria after an eight-month siege and swore to take revenge by slaughtering the populace until the river of blood reached his horse's knees; the threat was mitigated when his mount stumbled as he rode into the city. In gratitude, the citizens of Alexandria erected a statue of the horse.

The only extended period during the turbulent 3rd century AD in which Egypt was lost to the central imperial authority was 270–272, when it fell into the hands of the ruling dynasty of the Syrian city of Palmyra. Fortunately for Rome, the military strength of Palmyra proved to be the major obstacle to the overrunning of the Eastern Empire by the powerful Sāsānian monarchy of Persia.

Internal threats to security were not uncommon but normally were dissipated without major damage to imperial control. These included rioting between Jews and Greeks in Alexandria in the reign of Caligula (Gaius Caesar Germanicus; ruled AD 37–41); a serious Jewish revolt under Trajan (ruled AD 98–117); a revolt in the Delta in AD 172 that was quelled by Avidius Cassius; and a revolt centred on the town of Coptos (Qift) in AD 293/294 that was put down by Galerius, Diocletian's imperial colleague.

Revolts against Rome

Administration and economy under Rome. The Romans introduced important changes in the administrative system, aimed at achieving a high level of efficiency and

maximizing revenue. The duties of the prefect of Egypt combined responsibility for military security through command of the legions and cohorts, for the organization of finance and taxation, and for the administration of justice. This involved a vast mass of detailed paperwork: one document of AD 211 notes that in a period of three days 1,804 petitions were handed into the prefect's office. But the prefect was assisted by a hierarchy of subordinate equestrian officials with expertise in particular areas. There were three or four *epistratēgoi* in charge of regional subdivisions; special officers were in charge of the emperors' private account, the administration of justice, religious institutions, and so on. Subordinate to them were the local officials in the nomes (*stratēgoi* and royal scribes) and finally the authorities in the towns and villages.

It was in these growing towns that the Romans made the most far-reaching changes in administration. They introduced colleges of magistrates and officials who were to be responsible for running the internal affairs of their own communities on a theoretically autonomous basis and, at the same time, were to guarantee the collection and payment of tax quotas to the central government. This was backed up by the development of a range of "liturgies," compulsory public services that were imposed on individuals according to rank and property to ensure the financing and upkeep of local facilities. These institutions were the Egyptian counterpart of the councils and magistrates that oversaw the Greek cities in the eastern Roman provinces. They had been ubiquitous in other Hellenistic kingdoms, but in Ptolemaic Egypt they had existed only in the so-called Greek cities (Alexandria, Ptolemais in Upper Egypt, Naukratis, and later Antinoöpolis, founded by Hadrian in AD 130). Alexandria lost the right to have a council, probably in the Ptolemaic period. When it recovered its right in AD 200 the privilege was diluted by being extended to the nome capitals (*mētropoleis*) as well. This extension of privilege represented an attempt to shift more of the burden and expense of administration onto the local propertied classes, but it was eventually to prove too heavy. The consequences were the impoverishment of many of the councillors and their families and serious problems in administration that led to an increasing degree of central government interference and, eventually, more direct control.

Taxation under Roman rule

The economic resources that this administration existed to exploit had not changed since the Ptolemaic period, but the development of a much more complex and sophisticated taxation system was a hallmark of Roman rule. Taxes in both cash and kind were assessed on land, and a bewildering variety of small taxes in cash, as well as customs dues and the like, was collected by appointed officials. A massive amount of Egypt's grain was shipped downriver both to feed the population of Alexandria and for export to Rome. Despite frequent complaints of oppression and extortion from the taxpayers, it is not obvious that official tax rates were very high. In fact the Roman government had actively encouraged the privatization of land and the increase of private enterprise in manufacture, commerce, and trade, and low tax rates favoured private owners and entrepreneurs. The poorer people gained their livelihood as tenants of state-owned land or of property belonging to the emperor or to wealthy private landlords, and they were relatively much more heavily burdened by rentals, which tended to remain at a fairly high level.

Overall, the degree of monetarization and complexity in the economy, even at the village level, was intense. Goods were moved around and exchanged through the medium of coin on a large scale and, in the towns and the larger villages, a high level of industrial and commercial activity developed in close conjunction with the exploitation of the predominant agricultural base. The volume of trade, both internal and external, reached its peak in the 1st and 2nd centuries AD. But by the end of the 3rd century AD, major problems were evident. A series of debasements of the imperial currency had undermined confidence in the coinage, and even the government itself was contributing to this by demanding more and more irregular tax payments in kind, which it channeled directly to the main consumers, the army personnel. Local administration by the councils was careless, recalcitrant, and inefficient; the evident need for firm and purposeful reform had to be squarely faced in the reigns of Diocletian and Constantine.

Society, religion, and culture. One of the more noticeable effects of Roman rule was the clearer tendency to classification and social control of the populace. Thus, despite many years of intermarriage between Greeks and Egyptians, lists drawn up in AD 4/5 established the right of certain families to class themselves as Greek by descent and to claim privileges attaching to their status as members of an urban aristocracy, known as the gymnasial class. Members of this group were entitled to lower rates of poll tax, subsidized or free distributions of food, and maintenance at the public expense when they grew old. If they or their descendants were upwardly mobile, they might gain Alexandrian citizenship, Roman citizenship, or even equestrian status, with correspondingly greater prestige and privileges. The preservation of such distinctions was implicit in the spread of Roman law and was reinforced by elaborate codes of social and fiscal regulations such as the "Rule-Book of the Emperors' Special Account." The "Rule-Book" prescribed conditions under which people of different status might marry, for instance, or bequeath property and fixed fines, confiscations, and other penalties for transgression. When an edict of the emperor Caracalla conferred Roman citizenship on practically all of the subjects of the empire in AD 212, the distinction between citizens and noncitizens became meaningless; but it was gradually replaced by an equally important distinction between *honestiores* and *humiliores* (meaning, roughly, upper and lower classes), groups that, among other distinctions, were subjected to different penalities in law.

Social and fiscal codes

Naturally, it was the Greek-speaking elite that continued to dictate the visibly dominant cultural pattern, though Egyptian culture was not moribund or insignificant; one proof of its continued survival can be seen in its reemergent importance in the context of Coptic Christianity in the Byzantine period. An important reminder of the mixing of the traditions comes from a family of Panopolis in the 4th century, whose members included both teachers of Greek oratory and priests in Egyptian cult. The towns and villages of the Nile Valley have preserved thousands of papyri that show what the literate Greeks were reading: the poems of Homer and the lyric poets, works of the classical Greek tragedians, and comedies of Menander, for example. The pervasiveness of the Greek literary tradition is strikingly demonstrated by evidence left by an obscure and anonymous clerk at the Fayyūm village of Karanis in the 2nd century AD. In copying out a long list of taxpayers, the clerk translated an Egyptian name in the list by an extremely rare Greek word that he could only have known from having read the Alexandrian Hellenistic poet Callimachus; he must have understood the etymology of the Egyptian name as well.

Alexandria continued to develop as a spectacularly beautiful city and to foster Greek culture and intellectual pursuits, though the great days of Ptolemaic court patronage of literary figures had passed. But the flourishing interest in philosophy, particularly Platonic, had important effects. The great Jewish philosopher and theologian of the 1st century, Philo of Alexandria, brought a training in Greek philosophy to bear on his commentaries on the Old Testament. This anticipates by a hundred years the period after the virtual annihilation of the great Jewish community of Alexandria in the revolt of AD 115–117, when the city was the intellectual crucible in which Christianity developed a theology that took it away from the influence of the Jewish exegetical tradition and toward that of Greek philosophical ideas. There the foundations were laid for the teaching of the heads of the Christian catechetical school, such as Clement of Alexandria. And in the 3rd century there was the vital textual and theological work of Origen, the greatest of the Christian Neoplatonists, without which there would hardly have been a coherent New Testament tradition at all.

Outside the Greek ambience of Alexandria, traditional Egyptian religious institutions continued to flourish in the towns and villages; but the temples were reduced to financial dependence on a state subvention (*syntaxis*) and they

became subject to stringent control by secular bureaucrats. Nevertheless, like the Ptolemies before them, Roman emperors appear in the traditional form as Egyptian kings on temple reliefs until the middle of the 3rd century; and five professional hieroglyph cutters were still employed at the town of Oxyrhynchus in the 2nd century. The animal cults continued to flourish, despite Augustus' famous sneer that he was accustomed to worship gods, not cattle. As late as the reign of Diocletian (AD 285–305) religious stelae preserved the fiction that in the cults of sacred bulls (best known at Memphis and at Hermonthis), the successor of a dead bull was "installed" by the monarch. Differences between cults of the Greek type and the native Egyptian cults were still very marked, in the temple architecture as in the status of the priests. Priests of Egyptian cult formed, in effect, a caste distinguished by their special clothing, whereas priestly offices in Greek cult were much more like magistracies and tended to be held by local magnates. Cult of Roman emperors, living and dead, became universal after 30 BC, but its impact is most clearly to be seen in the foundations of Caesarea (Temples of Caesar) and in religious institutions of Greek type, where divine emperors were associated with the resident deities.

Egyptian and Roman cults

One development that did have an important effect on this pagan religious amalgam, though it was not decisive until the 4th century, was the arrival of Christianity. The tradition of the foundation of the church of Alexandria by St. Mark cannot be substantiated, but a fragment of a text of the Gospel According to John provides concrete evidence of Christianity in the Nile Valley in the second quarter of the 2nd century AD. Inasmuch as Christianity remained illegal and subject to persecution until the early 4th century, Christians were reluctant to advertise themselves as such, and it is therefore difficult to know how numerous they were, especially because later pro-Christian sources may often be suspected of exaggerating the zeal and the numbers of the early Christian martyrs. But several papyri survive of the *libelli* submitted in the first official state-sponsored persecution of Christians, under the emperor Decius (ruled 249–251): these were certificates in which people swore that they had performed sacrifices to pagan gods in order to prove that they were not Christians. By the 290s, a decade or so before the great persecution of Diocletian, a list of buildings in the sizeable town of Oxyrhynchus, some 125 miles south of the apex of the delta, included two Christian churches, probably of the house-chapel type.

Arrival of Christianity

Egypt's role in the Byzantine Empire. Diocletian was the last reigning Roman emperor to visit Egypt, in AD 302. Within about 10 years of his visit, the persecution of Christians ceased. The end of persecution had such far-reaching effects that from this point on it is necessary to think of the history of Egypt in a very different framework. No single point can be identified as the watershed between the Roman and Byzantine periods, as the divide between the peace, culture, and prosperity of the Principate and the darker age of the Dominate, supposedly characterized by a more oppressive state machinery in the throes of decline and fall. The crucial changes occurred in the last decade of the 3rd century and the first three decades of the 4th. With the end of persecution of Christians came the restoration of the property of the church. In 313 a new system of calculating and collecting taxes was introduced, with 15-year tax cycles, called indictions, inaugurated retrospectively from the year 312. Many other important administrative changes had already taken place. In 296 the separation of the Egyptian coinage from that of the rest of the empire had come to an end when the Alexandrian mint stopped producing its tetradrachmas, which had been the basis of the closed currency system.

One other event that had an enormous effect on the political history of Egypt was the founding of Constantinople on May 11, 330. First, Constantinople was established as an imperial capital and an eastern counterpart to Rome itself, thus undermining Alexandria's traditional position as the first city of the Greek-speaking East. Second, it diverted the resources of Egypt away from Rome and the West. Henceforth, part of the surplus of the Egyptian grain supply, which was put at 8,000,000 *artabs* (about 300,000,000 litres) of wheat in an edict of the emperor Justinian of about 537 or 538, went to feed the growing population of Constantinople, and this created an important political and economic link. The cumulative effect of these changes was to knit Egypt more uniformly into the structure of the empire and to give it, once again, a central role in the political history of the Mediterranean world.

Founding of Constantinople and its effect on Egypt

The key to understanding the importance of Egypt in this period lies in seeing how the Christian Church came rapidly to dominate secular as well as religious institutions and to acquire a powerful interest and role in every political issue. The corollary of this was that the head of the Egyptian Church, the patriarch of Alexandria, became the most influential figure within Egypt, as well as the person who could give the Egyptian clergy a powerful voice in the councils of the Eastern Church. During the course of the 4th century, Egypt was divided for administrative purposes into a number of smaller units but the patriarchy was not, and its power thus far outweighed that of any local administrative official. Only the governors of groups of provinces (*vicarii* of dioceses) were equivalent, the praetorian prefects and emperors superior; and when a patriarch of Alexandria was given civil authority as well, as happened in the case of Cyrus, the last patriarch under Byzantine rule, the combination was very powerful indeed.

The turbulent history of Egypt in the Byzantine period can largely be understood in terms of the struggles of the successive (or, after AD 570, coexisting) patriarchs of Alexandria to maintain their position both within their patriarchy and outside it in relation to Constantinople. What linked Egypt and the rest of the Eastern Empire was the way in which the imperial authorities, when strong (as, for instance, in the reign of Justinian), tried to control the Egyptian Church from Constantinople, while at the same time assuring the capital's food supply and, as often as not, waging wars to keep their empire intact. Conversely, when weak they failed to control the church. For the patriarchs of Alexandria, it proved impossible to secure the approval of the imperial authorities in Constantinople and at the same time maintain the support of their power base in Egypt. The two made quite different demands, and the ultimate result was a social, political, and cultural gulf between Alexandria and the rest of Egypt, and between Hellenism and native Egyptian culture, which found a powerful new means of expression in Coptic Christianity. The gulf was made more emphatic after the Council of Chalcedon in 451 established the official doctrine that Christ was to be seen as existing in two natures, inseparably united. The council's decision in effect sent the Egyptian Coptic (now Coptic Orthodox) Church off on its own path of Monophysitism, which centred around a firm insistence on the singularity of the nature of Christ.

Power of the patriarchs of Alexandria

Despite the debilitating effect of internal quarrels between rival churchmen, and despite the threats posed by the hostile tribes of Blemmyes and Nubade in the south (until their conversion to Christianity in the mid-6th century), emperors of Byzantium still could be threatened by the strength of Egypt if it were properly harnessed. The last striking example is the case of the emperor Phocas, a tyrant who was brought down in 609 or 610. Nicetas, the general of the future emperor Heraclius, made for Alexandria from Cyrene, intending to use Egypt as his power base and cut off Constantinople's grain supply. By the spring of 610 Nicetas' struggle with Bonosus, the general of Phocas, was won, and the fall of the tyrant duly followed.

The difficulty of defending Egypt from a power base in Constantinople was forcefully illustrated during the last three decades of Byzantine rule. First, the old enemy, the Persians, advanced to the Nile Delta and captured Alexandria. Their occupation was completed early in 619 and continued until 628, when Persia and Byzantium agreed to a peace treaty and the Persians withdrew. This had been a decade of violent hostility to the Egyptian Coptic Christians; among other oppressive measures, the Persians are said to have refused to allow the normal ordination of bishops and to have massacred hundreds of monks in their cave monasteries. The Persian withdrawal hardly heralded the return of peace to Egypt.

In Arabia events were taking place that would soon

bring momentous changes for Egypt. These were triggered by the flight of the Prophet Muḥammad from Mecca to Medina and by his declaration in AD 632 of a holy Islāmic war against Byzantium. Ten years later, by Sept. 29, 642, the Arab general ʿAmr ibn al-ʿĀṣ was able to march into Alexandria, and the Arab conquest of Egypt, which had begun with an invasion three years earlier, ended in peaceful capitulation. The invasion itself had been preceded by several years of vicious persecution of Coptic Christians by the Chalcedonian patriarch of Alexandria, Cyrus, and it was he who is said to have betrayed Egypt to the forces of Islām.

Islāmic conquest of Egypt

The Islāmic conquest was not bloodless. There was desultory fighting at first in the eastern Delta, then the Fayyūm was lost in battle in 640, and a great battle took place at Heliopolis (now a suburb of Cairo) in July 640 in which 15,000 Arabs engaged 20,000 Egyptian defenders. The storming and capture of Trajan's old fortess at Babylon (on the site of the present-day quarter called Old Cairo) on April 6, 641, was crucial. By September 14 Cyrus, who had been recalled from Egypt 10 months earlier by the emperor Heraclius, was back with authority to conclude a peace. Byzantium signed Egypt away on Nov. 8, 641, with provision for an 11-month armistice to allow ratification of the treaty of surrender by the emperor and the caliph. In December 641 heavily laden ships were dispatched to carry Egypt's wealth to its new masters. Nine months later the last remnants of Byzantine forces had left Egypt in ships bound for Cyprus, Rhodes, and Constantinople, and ʿAmr ibn al-ʿĀṣ had taken Alexandria in the name of the caliph. The new domination by the theocratic Islāmic caliphate was more strikingly different than anything that had happened in Egypt since the arrival of Alexander the Great almost a thousand years earlier.

Byzantine government of Egypt. The reforms of the early 4th century had established the basis for another 250 years of comparative prosperity in Egypt, at a cost of perhaps greater rigidity and more oppressive state control. Egypt was subdivided for administrative purposes into a number of smaller provinces, and separate civil and military officials were established (the *praeses* and the *dux*). By the middle of the 6th century the emperor Justinian was eventually forced to recognize the failure of this policy and to combine civil and military power in the hands of the *dux* with a civil deputy (the *praeses*) as a counterweight to the power of the church authorities. All pretense of local autonomy had by then vanished. The presence of the soldiery was more noticeable, its power and influence more pervasive in the routine of town and village life. Taxes were perhaps not heavier than they had been earlier, but they were collected ruthlessly, and strong measures were sanctioned against those who tried to escape from their fiscal or legal obligations. The wealthier landowners probably enjoyed increased prosperity, especially as a result of the opportunity to buy state-owned land that had been sold into private ownership in the early 4th century. The great landlords were powerful enough to offer their peasant tenants a significant degree of collective fiscal protection against the agents of the state, the rapacious tax collector, the officious bureaucrat, or the brutal soldier. But, if the life of the average peasant did not change much, nevertheless the rich probably became richer, and the poor became poorer and more numerous as the moderate landholders were increasingly squeezed out of the picture.

The advance of Christianity. The advance of Christianity had just as profound an effect on the social and cultural fabric of Byzantine Egypt as on the political power structure. It brought to the surface the identity of the native Egyptians in the Coptic Church, which found a medium of expression in the development of the Coptic language—basically Egyptian written in Greek letters with the addition of a few characters. Coptic Christianity developed its own distinctive art too, much of it pervaded by the long-familiar motifs of Greek mythology. These motifs coexisted with representations of the Virgin and Child and with Christian parables and were expressed in decorative styles that owed a great deal to both Greek and Egyptian precedents. Although Christianity had made great inroads into the populace by AD 391, the year in which the practice of pagan religion was officially made illegal, it is hardly possible to quantify it or to trace a neat and uniform progression. It engulfed its pagan precedents slowly and untidily. In the first half of the 5th century a pagan literary revival occurred, centred on the town of Panopolis, and there is evidence that fanatical monks in the area attacked pagan temples and stole statues and magical texts. Outside the rarefied circles in which doctrinal disputes were discussed in philosophical terms, there was a great heterogeneous mass of commitment and belief. Both the Gnostics, who believed in redemption through knowledge, and the Manichaeans, followers of the Persian prophet Mani, for example, clearly thought of themselves as Christians. In the 4th century a Christian community, the library of which was discovered at Najʿ Ḥammādī in 1945, was reading both canonical and apocryphal gospels as well as mystical revelatory tracts. At the lower levels of society pagan magical practices remained ubiquitous and were simply converted into a Christian context.

The Egyptian Coptic Church

By the middle of the 5th century Egypt's landscape was dominated by the great churches, such as the magnificent Church of St. Menas (Abū Mīna), south of Alexandria, and by the monasteries. The latter were Egypt's distinctive contribution to the development of Christianity and were particularly important as strongholds of native loyalty to the Monophysite Church. The origins of Antonian communities, named for the founding father of monasticism, St. Anthony of Egypt (*c.* 251–356), lay in the desire of individuals to congregate about the person of a celebrated ascetic in a desert location, building their own cells, adding a church and a refectory, and raising towers and walls to enclose the unit. Other monasteries, called Pachomian after Pachomius, the founder of cenobitic monasticism, were planned from the start as walled complexes with communal facilities. The provision of water cisterns, kitchens, bakeries, oil presses, workshops, stables, and cemeteries and the ownership and cultivation of land in the vicinity made these communities self-sufficient to a high degree, offering their residents peace and protection against the oppression of the tax collector and the brutality of the soldier. But it does not follow that they were divorced from contact with nearby towns and villages. Indeed, many monastics were important local figures and many monastery churches were probably open to the local public for worship.

Monasticism in Egypt

The economic and social power of the Christian Church in the Nile Valley and Delta is the outstanding development of the 5th and 6th centuries. By the time of the Arab invasion, in the mid-7th century, the uncomplicated propaganda of Islām might have seemed attractive and drawn attention to the political and religious rifts that successive and rival patriarchs of the Christian Church had so violently created and exploited. But the advent of Arab rule did not suppress Christianity in Egypt. Some areas remained heavily Christian for several centuries more.

(A.K.B.)

FROM THE ISLĀMIC CONQUEST TO 1250

Medieval Egyptian history opens and closes with foreign conquests of Egypt: the Arab invasion led by ʿAmr ibn al-ʿĀṣ in 639 and the Napoleonic expedition of 1798 mark the beginning and end of an era. Within the context of Egyptian internal history alone, this era was one in which Egypt cast off the heritage of the past to embrace a new language and a new religion—in other words, a new culture. While it is true that the past was by no means immediately and completely abandoned and that many aspects of Egyptian life, especially rural life, continued virtually unchanged, it is nevertheless clear that the civilization of Islāmic Egypt diverged sharply from that of the Greco-Roman period and was transformed under the impact of Western occupation. The history of medieval Egypt is therefore largely a study of the processes by which Egyptian Islāmic civilization evolved, particularly the processes of Arabization and Islāmization. But to confine Egyptian history to internal developments is to distort it, for during the entire medieval period Egypt was a part of a great world empire; and within this broader context, Egypt's history is a record of its long struggle to dominate

an empire—a struggle that is not without its parallels, of course, in both ancient and modern times.

Period of Arab and Turkish governors (639–868). The sending of a military expedition to Egypt from the caliphal capital in Medina came in a second phase of the first Arab conquests. Theretofore the conquests had been directed against lands on the northern borders of Arabia and were in the nature of raids for plunder; they had grown in scale and momentum as the Byzantines and Persians put up organized resistance. By 635 the Arabs had realized that in order to meet this resistance effectively they must begin the systematic occupation of enemy territory, especially Syria, where the Byzantine army was determined to halt the Arab forays.

The Arab expedition to Egypt

The Arab conquest. The Arabs defeated the Byzantines and occupied the key cities of Syria and Palestine, and they vanquished the Persian army on the eastern front in Mesopotamia and Iraq. The next obvious step was to secure Syria against a possible attack launched from the Byzantine province of Egypt. Beyond this strategic consideration, Arab historians call attention to the fact that ʿAmr ibn al-ʿĀṣ, the Arab general who later conquered Egypt, had visited Alexandria as a youth and had himself witnessed Egypt's enormous wealth. In spite of the obvious economic gain to be had from conquering Egypt, the caliph ʿUmar, according to some sources, showed reluctance to detach ʿAmr's expedition from the Syrian army and even tried to recall the mission once it had embarked; but ʿAmr, with or without the Caliph's permission, undertook the invasion in 639 with a small army of some 4,000 men (later reinforced). With what seems astonishing speed the Byzantine forces were routed and had withdrawn from Egypt by 642. An attempt by a Byzantine fleet and army to reconquer Alexandria in 645 was quickly defeated by the Arabs.

Various explanations have been given for the speed with which the conquest was achieved, most of which stress the weakness of Byzantine resistance rather than Arab strength. Certainly the division of the Byzantine government and army into autonomous provincial units militated against the possibility of a concerted and coordinated response. Although there is only dubious evidence for the claim that the Copts welcomed the Arab invasion in the belief that Muslim religious tolerance would be preferable to Byzantine enforced orthodoxy and repression, Coptic support for their Byzantine oppressors was probably unenthusiastic at best.

Early Arab rule. In Egypt—as in Syria, Iraq, and Iran—the Arab conquerors did little in the beginning to disturb the status quo; as a small religious and ethnic minority, they thus hoped to make the occupation permanent.

Arab policies

Treaties concluded between ʿAmr and the *muqawqis* (presumably a title referring to Cyrus, archbishop of Alexandria) granted protection to the native population in exchange for the payment of tribute. There was no attempt to force, or even to persuade, the Egyptians to convert to Islām; the Arabs even pledged to preserve the Christian churches. The Byzantine system of taxation, combining a tax on land with a poll tax, was maintained, though it was streamlined and centralized for the sake of efficiency. The tax was administered by Copts, who staffed the tax bureau at all but the highest levels.

To the mass of inhabitants, the conquest must have made little practical difference, because the Muslim rulers left them alone, in the beginning at least, as long as they paid their taxes; if anything, their lot may have been slightly easier, because Byzantine religious persecution had ended. Moreover, the Arabs deliberately isolated themselves from the native population, according to ʿUmar's decree that no Arab could own land outside the Arabian Peninsula; this policy aimed at preventing the Arab tribal armies from dispersing and at ensuring a steady revenue from agriculture, on the assumption that the former landowners would make better farmers than would the Arab nomads.

As was their policy elsewhere, the conquerors refrained from using an established city such as Alexandria as their capital; instead, they founded a new garrison town laid out in tribal quarters. As the site for this town they chose the strategic apex of the triangle formed by the Nile Delta—at that time occupied by the Byzantine fortified township of Babylon. They named the town Fusṭāṭ, which is probably an Arabized form of the Greek term for "encampment" and gives a good indication of the nature of the earliest settlement. Like garrison towns founded by the Arabs in Iraq—Basra and Kūfah—Fusṭāṭ became the main agency of Arabization in Egypt inasmuch as it was the only town with an Arab majority and therefore required an extensive knowledge of Arabic from the native inhabitants.

The process of Arabization, however, was slow and gradual. Arabic did not displace Greek as the official language of state until 706, and there is evidence that Coptic continued to be used as a spoken language in Fusṭāṭ. Given the lack of pressure from the conquerors, the spread of their religion must have been even slower than that of their language. A mosque was built in Fusṭāṭ bearing the name of ʿAmr ibn al-ʿĀṣ, and each quarter of the town had its own smaller mosque. ʿAmr's mosque served not only as the religious centre of the town but also as the seat of certain administrative and judicial activities as well.

Although Alexandria was maintained as a port city, Fusṭāṭ, being built on the Nile bank, was itself an important port and remained so until the 14th century. ʿAmr enhanced the port's commercial significance by clearing and reopening Trajan's Canal, so that shipments of grain destined for Arabia could be sent from Fusṭāṭ to the Red Sea by ship rather than by caravan.

Egypt under the caliphate. For more than 200 years—that is, throughout the Umayyad caliphate and well into the ʿAbbāsid—Egypt was ruled by governors appointed by the caliphs. As a province in an empire, Egypt's status was much the same as it had been for centuries under foreign rulers whose main interest was to supply the central government with Egyptian taxes and grain. In spite of evidence that the Arab governors tried in general to collect the taxes equitably, taking into account the capacities of individual landowners to pay and the annual variations in agricultural yield, resistance to paying the taxes increased in the 8th century and sometimes erupted into rebellion in times of economic distress. Periodically, religious unrest was manifested in the form of political insurrections, especially in those exceptional times when a governor openly discriminated against the Copts by forcing them to wear distinctive clothing or, worse, by destroying their icons. Still, the official policy, especially in Umayyad times, was tolerance, partly for fiscal reasons. In order to maintain the higher tax revenues collected from non-Muslims, the Arab governors discouraged conversion to Islām and even required those who did convert to continue paying the non-Muslim tax. New Christian churches were sometimes built, and the government took an interest in the selection of patriarchs.

Resistance to taxation

More than just a source of grain and taxes, Egypt also became a base for Arab-Muslim expansion, by both land and sea. The former Byzantine shipyards in Alexandria provided the nucleus of the Egyptian navy, which between 649 and 669 joined in expeditions with the Syrian navy against Rhodes, Cyprus, and Sicily and defeated the Byzantine navy in a major battle at Phoenix in 655. By land, the Arab armies advanced both to the south and to the west. As early as 651–652 the governor of Egypt invaded Nubia and imposed a treaty that required the Nubians to pay an annual tribute and to permit the unmolested practice of Islām in the province. Raids against North Africa by Arab armies based in Egypt began in 647; by 670 the Arabs had succeeded in establishing a garrison city in Ifrīqīyah (now Tunisia), called al-Qayrawān (Kairouan), which thenceforth displaced Egypt as the base for further expansion.

While some Arabs were passing through Egypt on their way to campaign in North Africa, others were being sent to the Nile Valley on a permanent basis. In addition to tribal contingents that at times escorted newly appointed governors to Egypt (some of which settled in towns), tribesmen were sometimes imported and settled in an effort to increase the Arab-Muslim concentration in the vicinity of Fusṭāṭ. The settlement of large numbers of anarchic tribesmen in Egypt, with tribal ties and allegiances elsewhere in the empire, meant that Egypt became embroiled in po-

Civil strife

litical difficulties with the central government. Civil strife centring around the assassination of the caliph ʿUthmān (656) began in Egypt, where the tribesmen resented the favouritism shown by the caliph to members of his own family. Uprisings led by the dissident Khārijite sect (the Seceders) were frequent in the mid-8th century. In the 9th century the caliph Maʾmun himself led an army from Iraq to put down a rebellion raised both by tribesmen and by Copts; repression of the Copts accompanying their defeat in 829–830 is usually cited as an important factor in accelerating conversion to Islām.

The difficulty inherent in ruling Egypt from Baghdad, which was itself undergoing stress and turbulence, is evident from the rapid turnover in governors assigned to Egypt; the ʿAbbāsid caliph Hārūn ar-Rashīd (ruled 786–809), for example, appointed 24 governors in a reign of 23 years. Possibly as a means of both removing the governorship from the level of tribal strife and paying the central government's Turkish troops, the caliphs began assigning Egypt to Turks rather than to Arabs. But this policy resulted in no tangible improvement in the administration of Egyptian affairs until 868, when the reign of Aḥmad ibn Ṭūlūn inaugurated a new phase of medieval Egyptian history.

The Ṭūlūnid dynasty (868–905). Though short-lived, the Ṭūlūnid dynasty succeeded in restoring a measure of Egypt's ancient glory. For the first time since the pharaohs, Egypt became virtually autonomous and the bulk of its revenues remained within its borders. What is more, Egypt became the centre of a small empire when Ibn Ṭūlūn conquered Syria in 878–879. These developments were paralleled in other provinces of the ʿAbbāsid Empire and were the direct result of the decline of the caliph's power. In order to strengthen their armies, the ʿAbbāsid caliphs had begun early in the 9th century to form contingents of Turkish slaves. To finance these new military formations and, in particular, to pay the Turkish commanders who headed them, the caliphs began to give them administrative grants (*iqṭāʿ* in Arabic, usually translated "fief") consisting of tax revenues from certain territories. In 868 Egypt was granted as a fief to the Turkish general Babak, who chose to remain in Iraq but appointed his stepson, Aḥmad ibn Ṭūlūn, as his agent in Egypt. Ibn Ṭūlūn's great achievement was that he quickly established his own authority in Egypt and backed it up with an army of his own creation, powerful enough to defy the central government of Baghdad and to embark upon foreign expansion.

Ibn Ṭūlūn's first step was to eliminate possible rivals in Egypt. From an early date the administration of Egypt had been divided between the *amīr* (military governor), appointed by the caliph, and the *ʿāmil* (fiscal officer), who was sometimes appointed by the caliph, sometimes by the governor. When Ibn Ṭūlūn entered Egypt in 868 he found the office of *ʿāmil* filled by one Ibn al-Mudabbir, who over a period of years had gained control of Egyptian finances, enriching himself in the process, and was therefore reluctant to acknowledge Ibn Ṭūlūn's authority. A struggle for power soon broke out between the two, which ended four years later with the transfer of Ibn al-Mudabbir to Syria and the assumption of his duties and powers by Ibn Ṭūlūn. An even more important step was the acquisition of an army that would be independent of the caliphate and loyal to Ibn Ṭūlūn. To build such an army, Ibn Ṭūlūn resorted to the same method the caliphs themselves used—the purchase of slaves who could be trained as military units loyal to their owner.

In 877, when Ibn Ṭūlūn failed to pay Egypt's full contribution to the ʿAbbāsid campaign against a black slave uprising in Iraq, the caliphal government, dominated by the caliph's brother al-Muwaffaq, realized that Egypt was slipping from imperial control. An expedition dispatched by al-Muwaffaq to remove Ibn Ṭūlūn from the governorship failed. Taking advantage of the caliphate's preoccupation with the revolt, Ibn Ṭūlūn in 878 invaded Syria, where he occupied the principal cities and garrisoned them with his troops. Thereafter he signified his autonomy by imprinting his name on the coinage along with the name of the caliph. Although the regent al-Muwaffaq lacked the resources to engage Ibn Ṭūlūn in battle, he did have him publicly cursed in the mosques of the empire as a means of retaliation.

Autonomy under Ibn Ṭūlūn

Internally, Ibn Ṭūlūn took active measures to raise Egyptian agricultural productivity and thereby to increase tax revenues; the huge surplus he left in the state treasury at his death in 884 is a measure of his success. Another tangible indication of his achievement for Egypt is an enormous mosque (still standing) that he erected in a suburb of Fusṭāṭ; in contrast, no building comparable in grandeur had even been contemplated by the governors who preceded him.

The great benefits Ibn Ṭūlūn had gained for Egypt by using its resources within the country were squandered by his son and successor, Khumārawayh. He expended huge sums on luxurious appointments for his residence and paid a fortune as a dowry for a daughter he married to the caliph al-Muʿtaḍid in 895. Nevertheless, Khumārawayh was able to maintain the Egyptian armies in the field, and he led them to victory both in Syria and in Mesopotamia. He resolved his father's conflict with the caliphate by a combination of arms and diplomacy, so that Khumārawayh's authority over Egypt, Syria, and Mesopotamia was given official caliphal recognition. This apparent strength evaporated when Khumārawayh was murdered in 896, leaving no funds with which his heir, a 14-year-old youth, could pay the troops. Both Egypt and Syria fell into anarchy, which lasted until 905 when a caliphal army invaded Egypt and momentarily restored it to the status of a province ruled by governors sent from Baghdad.

The Ikhshīdid dynasty (935–969). For 30 years the governors were unable to restore stability in Egypt. During this time, Egypt was subjected to attacks from the Fāṭimid state based in North Africa and to the rampages of an unruly domestic army. The appointment of Muḥammad ibn Ṭughj, from Sogdiana in Central Asia, as governor in 935 led to a repetition of Ibn Ṭūlūn's achievement; by bold measures Muḥammad established his authority over the treasury and the army, reasserted Egyptian influence in Syria, and won the governorship of the Holy Cities of Arabia (Mecca and Medina). In addition, he founded a dynasty; his sons inherited his Sogdian princely title of *ikhshīd,* but their authority was usurped by their Abyssinian slave tutor, Kāfūr, who ruled Egypt with the caliph's sanction. When Kāfūr died in 968 the Ikhshīdids were unable to maintain order in the army and the bureaucracy. In the following year the Fāṭimids took advantage of the disorder in Egypt to launch yet another attack, this one so successful that it led to the occupation of the country by a Berber army led by the Fāṭimid general Jawhar.

The Fāṭimid dynasty (969–1171). The establishment of the Fāṭimid caliphate in 973 in the newly built palace city of Cairo had dramatic consequences for the evolution of Islāmic Egypt. Politically, the Fāṭimids went a step further than the Ṭūlūnids by setting up Egypt as an independent rival to the ʿAbbāsid caliphate. In fact, an avowed aim of the early Fāṭimid propagandists was to achieve world dominion, eradicating the ʿAbbāsid caliphate in the process. For a variety of reasons they achieved neither of these goals; nevertheless, at the height of Fāṭimid power at the beginning of the 11th century, the Fāṭimid caliph could claim sovereignty over the whole North African coastal region, Sicily, the Hejaz and Yemen in Arabia, and southern Syria. Although actual political-military control was never firm except in Egypt, allegiance paid to the Fāṭimids by their provinces was just as meaningful as that paid to the ʿAbbāsids and for a time was certainly more widespread. Even when the Fāṭimid state fell into decline later in the 11th century and abandoned its imperial vision, Egypt continued to play an independent role in the Islāmic world under the leadership of Armenian generals who had gained control of the Fāṭimid armies.

Egypt set up as a rival to the ʿAbbāsid caliphate

Islāmization. It is difficult to estimate the religious change effected by the new dynasty except on the level of the governmental elite, which espoused the official doctrine of Ismāʿīlī Shīʿism—the branch that held all authority to inhere in the line of Ismāʿīl, who had predeceased both his father, the sixth ʿAlid *imām* Jaʿfar ibn Muḥammad, and his own son, Muḥammad at-Tamm. Because they believed that the Fāṭimid caliph was the only legitimate

leader, the practice of Sunnī (orthodox) Islām was theoretically outlawed in Fāṭimid domains. But the practical difficulties which the Ismāʿīlī minority faced in imposing its will on the Sunnī majority meant that the Muslim population of Egypt remained predominantly Sunnī throughout the Fāṭimid period. Certainly there was no public outcry when Saladin, who founded the Ayyūbid dynasty, restored Egypt to Sunnī rule in 1171. Regarding non-Muslims, the Fāṭimids, with one notable exception, were known for their tolerance, and the Copts continued to serve in the bureaucracy. Several Copts held the highest administrative post—the vizierate—without changing their religion. Jews also figured prominently in the government; in fact, a Jewish convert to Islām, Ibn Killis, was the first Fāṭimid vizier and is credited with laying the foundations of the Fāṭimid administrative system. Christians and Jews even managed to survive the reign of the mad caliph al-Ḥākim (ruled 996–1021), who ordered the destruction of Christian churches in Fāṭimid territory, including the Church of the Holy Sepulchre in Jerusalem, and offered his non-Muslim subjects the choice of conversion to Islām or expulsion from Fāṭimid territory. This period of persecution undoubtedly accelerated the rate of conversion to Islām, if only on a temporary and superficial level.

Lack of literature and learning

In comparison with Iraq, Egypt contributed relatively little to Arabic literature and Islāmic learning during the early ʿAbbāsid period. But the Fāṭimids' intense interest in propagating Ismāʿīlī Shīʿism through a network of missionary propagandists made Egypt an important religious and intellectual centre. The foundation of the mosque-college of al-Azhar as well as of other academies drew Shīʿite scholars to Egypt from all over the Muslim world and stimulated the production of original contributions in literature, philosophy, and the Islāmic sciences.

Arabization. The Arabization of Egypt continued at a gradual pace. The early Fāṭimids' reliance on Berber troops was soon balanced by the importation of Turkish, Sudanese, and Arab contingents. The Fāṭimids are said to have used thousands of nomadic Arabs in the Egyptian cavalry and to have further stimulated Arabization by settling large numbers of Arabian tribesmen in Upper Egypt to deprive the Qarmaṭians—their Ismāʿīlī enemies in Iraq and Arabia—of Arab tribal support. On the other hand, the Fāṭimids reduced the Arab population of Egypt in the mid-11th century, when they incited the Banū Hilāl and the Banū Ṣulaym tribes to emigrate from Egypt into the neighbouring Berber kingdom of Ifrīqīyah.

Growth of trade. One of the most far-reaching changes in Fāṭimid times was the growth of Egyptian commerce, especially in Fusṭāṭ, which had become the port city for Cairo, the Fāṭimid capital. Theretofore, Iraq in the east and Tunisia in the west had been flourishing centres for trade conducted both within the Muslim world and between the Muslim and the Christian empires of the West. A number of factors contributed to alter this situation in favour of Egypt. As centralized power declined in Iraq, Mesopotamia, and Syria during the 9th and 10th centuries, traffic on the trade routes across these areas also declined. In Egypt, however, the establishment of a strong government, which soon controlled the Red Sea and maintained a strong navy in the eastern Mediterranean, offered an attractive alternative for the international transit trade between the Orient and Christendom. In addition to having the political stability essential for trade, the Fāṭimids encouraged commerce by their low tariff policy and their noninterference in the affairs of merchants who did business in Egypt. These factors, along with increased European mercantile activity in the Italian cities, helped restore Egypt as a great international entrepôt.

Egyptian commerce

The end of the Fāṭimid dynasty. The Fāṭimid achievement in restoring to Egypt a measure of its ancient glory was remarkable but brief. Halfway through their history the political-religious authority of the Fāṭimid caliphs was vitiated by military uprisings that could be put down only by force. By 1163 the Fāṭimid caliph had been shunted aside in a power struggle between the vizier and the chamberlain, who were themselves so impotent that they had to seek help from the Sunnī and even from the crusader powers of Syria and Palestine. Thus began a series of invasions at the behest of Fāṭimid officials, which ended in 1169 with the occupation of Egypt by an army from Syria, one of whose commanders—Saladin—was appointed Fāṭimid vizier. Two years later Saladin restored Egypt to ʿAbbāsid allegiance, abolished the Fāṭimid caliphate, and, in effect, established the Ayyūbid dynasty.

The Ayyūbid dynasty (1171–1250). Under Saladin and his descendants, Egypt was reintegrated into the Sunnī world of the Eastern caliphate. Indeed, Egypt became champion of that world against the crusaders and, as such, chief target of the crusader armies. But this was a gradual process that required Saladin first to build an army strong enough to establish his power in Egypt and then to unite the factions of Syria and Mesopotamia under his leadership against the Franks. By so doing he reconstituted the Egyptian empire, which included, in addition to the areas just named, Yemen, the Hejaz, and, with the fall of Jerusalem (1187), a major part of the Holy Land.

Return to the Eastern caliphate under Saladin

The abolition of the Fāṭimid caliphate and the official reinstitution of Sunnī Islām seems to have caused little perturbation in Egypt except for an uprising by the Fāṭimid palace guard, quickly suppressed. This undoubtedly means that Ismāʿīlī Shīʿism was confined to Fāṭimid ruling circles.

Saladin's policies. Saladin's remission of all taxes not explicitly sanctioned by Islāmic law must have contributed to his own popularity as well as to the stability of his regime. To ensure the defense of his state against both internal and external enemies, he strengthened the fortifications of Cairo by building a citadel and extending the Fāṭimid city walls. Despite the major military and propagandistic efforts mounted against the crusaders, Saladin continued to treat the Christians of Egypt with tolerance; the Coptic Church thrived under the Ayyūbids, and Copts still served the government. Saladin also treated the Christians of Jerusalem with magnanimity after the conquest of that city.

Much to the consternation of the popes, trade between Egypt and the Italian city-states remained brisk, and the Egyptians were able to use raw materials provided by the Italian merchants to forge weapons against the crusaders. The administration of Egypt stayed in the hands of the vast, mainly civilian, bureaucracy, but was supervised by military officials.

Power struggles. The Ayyūbids introduced a significant change in the governance of their empire that was decisive for the history of their rule in Egypt. Though the Ayyūbids were themselves of Kurdish descent, Saladin followed the Turkish practice of assigning the provinces as fiefdoms to members of his family. In theory, such a measure would ensure the loyalty of the provinces to the central government of Egypt through the loyalty of Ayyūbid kinsmen to their family leader. In practice, however, the measure led to recurrent power struggles in which each governor used his province as a base from which to defy the supreme Ayyūbid power of Egypt. The sultans al-Malik al-ʿĀdil (died 1218) and al-Malik al-Kāmil (died 1238) each succeeded in reuniting Syria and Egypt under his own leadership. Kāmil, especially, was able to exploit Frankish attacks—in the form of the Fifth Crusade, directed against Damietta—to rally family and provincial support for the defense of Egypt. Nevertheless, given the dissension within the Ayyūbid empire, it was clearly in the interest of the Egyptian sultan to reach a peaceful settlement with the crusaders; this was achieved in 1229 by a truce between Kāmil and the Holy Roman emperor Frederick II. The agreement stipulated that Kāmil exchange possession of Jerusalem and other territory in the Holy Land for Frederick's guarantee to support the sultan against aggression from any source.

Growth of Mamlūk armies. The only real security for Ayyūbid Egypt lay in its independent military strength. This explains why one of the last sultans, al-Malik aṣ-Ṣāliḥ Ayyūb (died 1249), resorted to increased purchase of Turkish slaves—called Mamlūks, a name derived from the Arabic word for slave—as a means of manning his armies. Although slave troops had formed an important part of Egyptian armies since the time of Aḥmad ibn Ṭūlūn, their strength had been checked by racial dissension

among the various slave units and by the presence of nonslave elements. But after the death of aṣ-Ṣāliḥ Ayyūb in the course of the Sixth Crusade, which the Egyptians defeated thanks to the Mamlūk corps, the Mamlūks were able to exploit a palace feud and to elevate a member of their own ranks to the sultanate. Thus was established the Mamlūk sultanate, which lasted for two and a half centuries and brought Egypt to the peak of its evolution in the medieval period.

THE MAMLŪK AND OTTOMAN PERIODS (1250–1800)

The Mamlūk dynasty (1250–1517). During the Mamlūk period Egypt became the unrivaled political, economic, and cultural centre of the eastern Arabic-speaking zone of the Muslim world. Symbolic of this development was the reestablishment in 1261 under the Mamlūks of the 'Abbāsid caliphate in Cairo (the Mongols had abolished the caliphate when they invaded Baghdad in 1258). Although the caliph enjoyed little authority and had no power, the mere fact that the Mamlūks chose to maintain the institution in Cairo is a measure of their determination to dominate Arabic Islām. It is curious that the Mamlūks—all of whom were of non-Arab, non-Muslim origin and some of whom knew little if any Arabic—established a regime that saved a substantial portion of Muslim territory from pagan domination and established Egypt's supremacy in Arabic culture.

Egypt as the centre of eastern Islām under the Mamlūks

Political life. The political history of the Mamlūk state is complex; during their 264-year reign, no fewer than 45 Mamlūks gained the sultanate, and once, in desperate circumstances, a caliph (in 1412) was briefly installed as sultan. At times individual Mamlūks succeeded in establishing dynasties, most notably Sultan Qalā'ūn (ruled 1279–90), whose progeny ruled Egypt, with two short interruptions, until 1382. Often the Mamlūks chose to allow a sultan's son to succeed his father only for as long as it took another Mamlūk to build up enough support to seize the throne for himself. In reality there was no principle of legitimacy other than force, for without sufficient military power a sultan could expect to be overthrown by a stronger Mamlūk. (D.P.L.)

Nevertheless, several sultans succeeded in harnessing the energies of the Mamlūk system to establish internal stability and to embark on foreign conquests. Soon after the Mamlūk victory over the Mongols at 'Ayn Jālūt in 1260, Baybars I seized power. He was the true founder of the Mamlūk state, and he compaigned actively and with success against the remaining crusader possessions in Palestine and Syria. He ruled until 1277. During the long reign of al-Malik an-Nāṣir (ruled 1293–1341), the Mamlūks concluded a truce with the Mongols (1323) after several major battles and, despite widespread famine, outbreaks of religious strife, and Bedouin uprisings, maintained economic prosperity in Egypt and peaceful relations with foreign powers, both Muslim and Christian. (D.P.L./D.S.Ri.)

Although the state began to decline politically and economically after the death of Nāṣir in 1341, Egypt continued to dominate Eastern Arabdom. But the cumulative effect of the plague, which swept Egypt in 1348 and on many occasions subsequently; Timur's victory in Syria in 1400; and Egypt's loss to the Portuguese of control over the Indian trade, along with the sultans' inability to keep their refractory Mamlūk corps under control, gradually sapped the strength of the state. The best efforts of such a vigorous sultan as Qā'it Bāy (ruled 1468–96) failed to make Egypt strong enough to defend its Syrian empire against raids by the Turkoman states of Anatolia and Azerbaijan and campaigns of the Ottoman Turks.

Contributions to Arabic culture. By the time of the Mamlūks, the Arabization of Egypt must have been almost complete. Arabic had been the language of the bureaucracy since the early 8th century and the language of religion and culture even longer. Moreover, the prevalence of Arabic as a written and spoken language is attested by the discovery in the geniza (storeroom) of a Cairo synagogue of thousands of letters and documents—called the "Geniza Documents"—dating from the 11th through the 13th century. Though often written in Hebrew characters, the actual language of most of these documents is Arabic, which proves that Arabic was widely used even by non-Muslims. The main incentive for learning Arabic must have come from the desire of a subject population to learn the language of the ruling elite. The immigration of Arab tribesmen during the early centuries of the occupation, and their intermarriage with the indigenous inhabitants, must also have contributed to the gradual spread of Arabic in Egypt.

Prevalence of Arabic in Mamlūk Egypt

The specific Mamlūk contribution to Arabic culture, however, lay above all in the military achievement. By defeating the Mongols, the Mamlūks provided a haven in Syria and in Egypt for Muslims fleeing from Mongol devastation. The extent of this haven was narrowed by subsequent Mongol attacks against Syria, one of which led to a brief Mongol occupation of Damascus in 1294–95, so that Egypt received an influx of refugees from Syria itself as well as from areas farther east.

This accidental displacement of scholars and artisans into Egypt does not, however, wholly account for the efflorescence of certain types of cultural activity under the Mamlūks. In the same way as they supported the caliphate as a visible symbol of their legitimate claim to rule Islāmic territory, the Mamlūks cultivated and patronized religious leaders whose skills they needed in administering their empire and in directing the religious sentiments of the masses into safe (*i.e.,* nondisruptive) channels. Those divines who cooperated with the state were rewarded with government offices, in the case of the *'ulamā'* (religious scholars), and with endowed monasteries, in the case of the Ṣūfīs (mystics). On the other hand, those who dared criticize the prevailing social and moral order were thrown into prison (such was the fate of the famous legist, Ibn Taymīyah, who, having emigrated from Mesopotamia in order to escape the Mongols, was incarcerated in Cairo by the Mamlūks and their religious functionaries for spreading seditious doctrines).

Concrete evidence of the stimulus the Mamlūks gave to cultural life can be found chiefly in the fields of architecture and historiography. Dozens of public buildings erected under Mamlūk patronage are still standing in Cairo and include mosques, colleges, hospitals, monasteries, and caravansaries. Historical writing under the Mamlūks was equally monumental, in the form of immense chronicles, biographical dictionaries, and encyclopaedias.

Religious life. The Mamlūk period is also important in Egyptian religious history. With few and therefore notable exceptions, the Muslim rulers of Egypt had seldom interfered with the lives of their Christian and Jewish subjects so long as these groups paid the special taxes levied on them in exchange for state protection. Indeed, both Copts and Jews had always served in the Muslim bureaucracy, sometimes in the very highest administrative positions. Even the Crusades apparently failed to upset the delicate balance between Muslims and Christians. Trade with the Italian city-states had certainly continued, and there is no evidence that the local Christians were held accountable for the crusader invasions of Egypt. While it is true that Saladin dismissed all Copts from the bureaucracy and imposed sumptuary laws on them, this policy was abandoned by his successors in their desire to reach an accommodation with the crusaders.

With the establishment of the Mamlūk dynasty, however, it is generally agreed that the lot of the Christians, both in Egypt and in Syria, took a distinct turn for the worse. One indication of this change is the increased production of anti-Christian polemics written by Muslim theologians. A possible reason for the change may have been the association of Christians with the Mongol peril. Because the Mongols used Christian auxiliaries in their armies—Georgians and Armenians in particular—they often spared the Christian populations of towns they conquered, while slaughtering the Muslims. Also, the diplomatic efforts aimed at uniting the Mongols with Christian European powers in a joint crusade against the Muslims might have contributed to the Mamlūks' distrust of the Christians. But the dissatisfaction seems to have originated not so much with the Mamlūk rulers as with the masses, and it seems to have been directed not so much against Chris-

Anti-Christian feelings

tians' sympathy for the Mongols as against their privileged position and role in the Mamlūk state.

On several occasions popular resentment against the Copts' conspicuous wealth and their employment in the government was manifested in public demonstrations. Both Muslims and Christians resorted to arson, burning the others' sanctuaries, to express their hatred. Under such pressure, the Mamlūk government dismissed Christians from the bureaucracy on no fewer than nine occasions between 1279 and 1447, and in 1301 it ordered all the churches in Egypt closed. As a result of these intermittent persecutions and the destruction of churches, it is believed that the rate of conversion to Islām accelerated markedly in the Mamlūk period and that Coptic virtually disappeared except as a liturgical language. By the end of the Mamlūk dynasty, the Muslims may well have reached the same numerical superiority that they enjoy in modern times—a ratio of more than 10 to one.

Economic life. In trade and commerce, the Mamlūk period marks the zenith of medieval Egyptian economic history. During the 13th and 14th centuries (as long, that is, as the sultanate was able to maintain order in Egypt), trade was heavy with Mediterranean and Black Sea ports and with India. The Oriental trade was controlled largely by a group of Muslim merchants known as the Kārimīs; the Mediterranean trade was left to European traders, whom the Mamlūks allowed certain privileges in Alexandria. By the 15th century, however, Egypt's commercial importance rapidly deteriorated as the result of population losses, increased government interference in commerce, Bedouin raiding, and Portuguese competition in the Indian trade.

The Ottomans (1517–1798). With the Ottomans' defeat of the Mamlūks in 1516–17, Egyptian medieval history had come full circle, as Egypt reverted to the status of a province governed from Istanbul. Again the country was exploited as a source of taxation for the benefit of an imperial government and as a base for foreign expansion. The economic decline that had begun under the late Mamlūks continued, and with it came a decline in Egyptian culture.

Economic decline under Mamlūks

Some historians attribute the lethargy of Ottoman Egypt solely to Ottoman domination. But although Ottoman policy was geared to imperial, not Egyptian, needs, it was obviously to the rulers' benefit to provide a stable government that would maintain Egyptian agriculture at a high level of productivity and would promote the transit trade. To a certain extent Ottoman actions served these purposes. The decisive factor that ultimately undermined Ottoman policies was the perpetuation of the former Mamlūk elite; though they collaborated with the Ottoman government, they often defied it and in the end they dominated it. By and large the history of Ottoman Egypt concerns the process by which the conquered Mamlūks reasserted their power within the Egyptian state.

The Ottoman conquest. From the conquest itself, the Ottoman presence in Egypt was entangled with Mamlūk factionalism. There is no doubt that the Ottomans invaded Syria in 1516 to break an incipient coalition against Ottoman expansion between the Ṣafavids of Persia and the Mamlūks of Egypt and Syria. The long-standing enmity between the Ottomans and the Mamlūks arose from their contest to control the Turkoman frontier states north of Syria. After the Ottomans strengthened their hold over eastern Anatolia in 1514, it was only natural that the Mamlūks should attempt to bolster their forces in northern Syria and exchange diplomatic missions with the Ṣafavids. The Ottoman sultan Selim the Grim responded by attacking the reinforced Mamlūk army in Syria, probably as a preliminary step in a new campaign against the Ṣafavids. In 1516, after Selim had defeated the Mamlūks at Marj Dābiq (north of Aleppo), Ottoman goals had probably been met, especially since the Mamlūk sultan Qānṣūh al-Ghaurī died in the battle. But the Mamlūks rallied around a new sultan in Cairo, who refused to accept Selim's terms for a settlement. Spurred on by the Mamlūk traitor Khair Bey, Selim marched against Egypt in 1517, defeated the Mamlūks, and installed Khair Bey as Ottoman governor. Khair Bey died in 1552; thereafter, the Ottoman viceroy (called *vali*), with the title of pasha, was sent from Istanbul.

Mamlūk factionalism

Ottoman administration. In 1525 the Ottoman administration of Egypt was defined and codified by the Ottoman grand vizier, Ibrahim Paşa, who was dispatched to Egypt for this purpose by the sultan Süleyman the Magnificent. According to the terms of Ibrahim Paşa's decree (*qanuṇname*), Egypt was to be ruled by a viceroy aided by an advisory council (divan) and an army comprising both Ottoman and local corps. The collection of taxes and the administration of the four provinces into which Egypt was divided were assigned to inspectors (*kashifs*). Although the Egyptian government was headed by bureaucratic officials sent from Istanbul, and supported by Ottoman troops, the Mamlūks were able to penetrate both the bureaucracy and the army. The *kashifs* were often drawn from Mamlūk ranks; three of the seven military corps formed by the Ottomans in the 16th century were recruited in Egypt, one of which—the Circassians—was composed of Circassian Mamlūks. Their service in the army enabled the Mamlūk amirs to secure high-ranking military posts that entitled them to serve on the divan itself.

By the 17th century a distinct elite bearing the title of bey had emerged, which consisted largely of Mamlūk amirs. These beys held no specific offices but were nevertheless paid a salary by the Ottoman government. The elite was perpetuated through the old Mamlūk system of purchasing slaves, giving them military training, then freeing them and attaching them to one of the great Mamlūk houses of Egypt. Thus, for all practical purposes, the Mamlūks maintained themselves as an elite throughout the Ottoman period. They were no longer the only political-military elite, as they had been in the past, but they ultimately succeeded in reestablishing their dominance. Yet the chief obstacle to the growth of their power was not so much the Ottoman ruling hierarchy as it was their own factionalism. During the 17th and 18th centuries the Mamlūks were divided into two great rival houses—the Faqariyya and the Qasimiyya—whose mutual hostility often broke out into fighting and impaired the strength of Mamlūks as a bloc.

The beys

Mamlūk power under the Ottomans. In spite of internal dissension and the resistance of the non-Mamlūk hierarchy, the Mamlūks had emerged by the early 18th century as the supreme power in Egyptian politics. While the beys continued to acknowledge the authority of the Ottoman viceroy and to send tribute to Istanbul, the strongest single figure in Egypt was the bey who held the newly coined title of *shaykh al-balad* ("chief of the city"), which signified that he was recognized by the other beys as their chief. The Mamlūks' rise to power was climaxed by the careers of two amirs—'Alī Bey and Abū Dhahab—both of whom secured from the Sublime Porte (Ottoman government) de facto recognition of their autonomy in Egypt (1768–76) and even undertook military campaigns in Syria and the Hejaz. The Ottomans attempted to end the Mamlūk domination by sending an army to Egypt in 1786. Although it was initially successful, this attempt failed and the troops were withdrawn a year later. A Mamlūk duumvirate was reestablished, and it lasted until Napoleon invaded Egypt in 1798.

Expansion. During the 16th century, when their regime in Egypt was strongest, the Ottomans used Egypt as a base for expansion to the south. Like the Mamlūk rulers before them, they attempted to control the southern approaches to Egypt by instituting their authority in Nubia; this they achieved by annexing Nubia as far south as the Third Cataract. Elsewhere, they undertook to reassert Egyptian command of the Red Sea, which the Portuguese had begun to contest during the early 16th century. Ottoman fleets and troops captured Yemen and Aden (1536–46) and thus dominated the lower Red Sea; in 1557 they strengthened this position by setting up a colony on the Abyssinian coast at Mitsiwa (Massawa). In the 17th century these outposts began to lose their importance as Ottoman and Portuguese power began to decline and the Dutch took over the spice trade.

Culture. Given the political instability and the economic decline that had prevailed in Egypt since late Mamlūk times, it is not surprising that the culture of Ottoman Egypt lacked vitality. Perhaps the most telling example of intellectual quiescence was the dramatic decline in the

Impact of political instability

quantity of historical works produced in Egypt. As already noted, the Mamlūk period is renowned for the number and quality of its historians, partly because the amirs patronized court historians; by contrast, in almost three centuries of Ottoman rule Egypt produced only one historian worthy of note, al-Jabartī (died 1825), famous for his observations on the French occupation. The Ottomans also fell short of the Mamlūks' achievement in architecture; there is no lack of public buildings erected under Ottoman patronage, but even the best of these are imitations of the Byzantine basilica, which had been adopted as the model for mosques.

Religious affairs. Like all previous Muslim governments, the Ottomans continued to employ Copts in the financial offices of the bureaucracy. The Ottomans allowed the caliphate, so assiduously preserved in its nominal form by the Mamlūks, to lapse. At first the caliph was installed in Istanbul by Selim the Grim. Later the caliph—the last of the ʿAbbāsid line—returned to Egypt, where he died in the reign of Süleyman. The claim that the caliph had transferred his authority to the Ottoman sultan is an 18th-century invention. (D.P.L.)

FROM THE FRENCH TO THE BRITISH OCCUPATION (1798–1882)

The French occupation and its consequences (1798–1805). Although several projects for a French occupation of Egypt had been advanced in the 17th and 18th centuries, the purpose of the expedition that sailed under Napoleon Bonaparte from Toulon in May 1798 was specifically connected with the war against Britain. Bonaparte had discounted the feasibility of an invasion of England but hoped, by occupying Egypt, to damage British trade, to threaten India, and to obtain assets for bargaining in any future peace settlement. Meanwhile, as a colony under the benevolent and progressive administration of Revolutionary France, Egypt would be regenerated and regain its ancient prosperity. The military and naval forces were therefore accompanied by a commission of scholars and scientists to investigate and report the past and present condition of the country.

Eluding the British Mediterranean fleet under Lord Nelson, the French landed at Abū Qīr (Aboukir) Bay on July 1 and took Alexandria the next day. In an Arabic proclamation, Bonaparte assured the Egyptians that he came as a friend to Islām and the Ottoman sultan, to punish the usurping Mamlūks and to liberate the people. From Alexandria the French advanced on Cairo, defeating Murad Bey at Shubrākhīt (July 13), and again decisively at Imbābah, opposite Cairo in the so-called Battle of the Pyramids on July 21. Murad fled to Upper Egypt, while his colleague, Ibrāhīm Bey, together with the Ottoman viceroy, made his way to Syria.

Battle of the Pyramids

After entering Cairo (July 25), Bonaparte sought to conciliate the population, especially the religious leaders (*ʿulamāʾ*), by demonstrating his sympathy with Islām and by establishing councils (divans) as a means of consulting Egyptian opinion. The destruction of the French fleet at Abū Qīr by Nelson in the so-called Battle of the Nile on August 1 virtually cut Bonaparte's communications and made it necessary for him to consolidate his rule and to make the expeditionary force as self-sufficient as possible. The savants, organized in the Institut d'Égypte, played their part in this. Meanwhile, Egyptian resentment at alien rule, administrative innovations, and the growing fiscal burden of military occupation was exacerbated when the Ottoman sultan, Selim III (1789–1807), declared war on France on September 11. An unforeseen revolt in Cairo on October 21 was suppressed after an artillery bombardment that ended any hopes of cordial Franco-Egyptian coexistence.

Ottoman Syria, dominated by Aḥmad al-Jazzār, the governor of Acre, was the base from which French-occupied Egypt might most easily be threatened, and Bonaparte resolved to deny it to his enemies. His invasion force crossed the frontier in February 1799 but failed to take Acre after a protracted siege (March 19–May 20), and Bonaparte evacuated Syrian territory. A seaborne Ottoman invading force landed at Abū Qīr in July but failed to maintain its bridgehead. At this point Bonaparte resolved to return to France and succeeded in slipping away on August 22, past the British fleet.

His successor as general in chief, Jean-Baptiste Kléber, viewed the situation of the expeditionary force with pessimism and, like many of the soldiers, wished to return to the theatre of war in Europe. He therefore entered into negotiations with the Ottomans and by the Convention of al-ʿArīsh (Jan. 24, 1800) agreed to evacuate Egypt. Sir Sydney Smith, the British naval commander in the eastern Mediterranean, sponsored the convention, but in this he had exceeded his powers and was instructed by his superior officer, Admiral Lord Keith, to require the French to surrender as prisoners of war. Although the Ottoman reoccupation was well underway, Kléber and the French determined on resistance and defeated the Turkish forces at the Battle of Heliopolis (March 20). A second revolt of Cairo, fomented by Ottoman fugitives, took about a month to suppress; but French authority had been restored when Kléber was assassinated by a Syrian Muslim, Sulaymān al-Ḥalabī, on June 14.

His successor, ʿAbd Allāh Jacques Menou, a French officer (and former nobleman) who had turned Muslim, was determined to maintain the occupation and administered at first a tolerably settled country, although he lacked the prestige of his two predecessors. In 1801 a threefold invasion of Egypt began. British troops were landed at Abū Qīr in March, while the Ottomans advanced from Syria. Shortly afterward, British Indian forces were landed at Quṣayr on the Red Sea coast. The French garrison in Cairo capitulated in June and Menou himself at Alexandria in September.

Surrender of the French

The brief episode of the French occupation was to be significant for Egypt in several ways. The arrival of a European army accompanied by scholars and scientists appropriately inaugurated the impact of the West, which was to be felt increasingly in the next 150 years. Egypt, protected for five centuries by the Mamlūk and Ottoman sultanates, was no longer immune from European attack: it had become an object of the contending policies of France and Britain, a part of the "Eastern Question." Bonaparte's savants had little success in interpreting Western culture to the traditionalist *ʿulamāʾ* of Cairo; their achievement was rather to unveil Egypt to Europe. They uncovered the celebrated Rosetta Stone, which held a trilingual inscription making it possible to decipher hieroglyphs and which thus laid the foundation of modern Egyptology. Their reports and monographs were collected in the monumental *Description de l'Égypte* ("Description of Egypt"), which was published in parts from 1809 to 1828 in Paris.

Of more immediate consequence for Egypt was the effect of the French occupation upon internal politics. The Mamlūk ascendancy was fatally weakened. Murad Bey, who had made his peace with the French, died shortly before their capitulation in 1801; and Ibrāhīm Bey, who returned to Egypt with the Ottomans, had henceforward little power. The new Mamlūk leaders, ʿUthmān Bey al-Bardīsī and Muḥammad Bey al-Alfī, former retainers of Murad, headed rival factions and had in any case to reckon with the British and Ottoman occupation forces. In March 1803 the British were evacuated in accordance with the Peace of Amiens. But the Ottomans, determined to reassert their control over Egypt, remained, establishing their power through a viceroy and an occupying army, in which the most effective fighting force was an Albanian contingent. The Albanians, however, acted as an independent party and in May 1803 mutinied and installed their own leader as acting viceroy. When he was assassinated shortly afterward, the command of the Albanians passed to his lieutenant, Muḥammad ʿAlī (born 1769), who, during the ensuing two years, cautiously strengthened his own position at the expense of both the Mamlūks and the Ottomans.

Ottomans remain in Egypt

Muḥammad ʿAlī and his successors (1805–82). In May 1805 a revolt broke out in Cairo against the Ottoman viceroy, Khūrshīd Pasha. The *ʿulamāʾ* invested Muḥammad ʿAlī as viceroy. For some weeks there was street fighting, and Khūrshīd was besieged in the Citadel. In July Sultan Selim III confirmed Muḥammad ʿAlī in office and the revolt ended.

Muḥammad ʿAlī's viceroyalty was marked by a series of military successes, some of which were attended by political failures that frustrated his wider aims. After the renewal of war between Britain and Napoleonic France in 1803, Egypt again became an area of strategic significance. A British expedition occupied Alexandria in 1807 but failed to capture Rosetta and, after a defeat at the hands of Muḥammad ʿAlī's forces, was allowed to withdraw.

Military expansion. In Arabia, the domination of Mecca and Medina by puritanical Wahhābī Muslims was a serious embarrassment to the Ottoman sultan, who was the titular overlord of the Arabian territory of the Hejaz and the leading Muslim sovereign. At the invitation of Sultan Mahmud II (1808–39), Muḥammad ʿAlī sent an expedition to Arabia that between 1811 and 1813 expelled the Wahhābīs from the Hejaz. In a further campaign (1816–18), Ibrāhīm Pasha, the viceroy's eldest son, defeated the Wahhābīs in their homeland of Najd and brought central Arabia within Egyptian control. In 1820–21 Muḥammad ʿAlī sent an expedition up the Nile and conquered much of what is now the northern Sudan. By so doing, he made himself master of one of the principal channels of the slave trade and began an African empire that was to be expanded under his successors.

After the outbreak of the Greek insurrection against Ottoman rule, Muḥammad ʿAlī, at Sultan Mahmud II's request, suppressed the Cretan revolt in 1822. In 1825 Ibrāhīm began a victorious campaign in the Morea in southern Greece, where his military success provoked intervention by the European powers and brought on the destruction of the Ottoman and Egyptian fleets at the Battle of Navarino in October 1827. The Morea was evacuated the following year.

Muḥammad ʿAlī's invasion of Syria

In 1831 Muḥammad ʿAlī embarked upon the invasion of Syria. His pretext was a quarrel with the governor of Acre, but deeper considerations were involved, particularly the growing strength of the Sultan, which might threaten his own autonomy. Syria, moreover, was strategically important; and its products, especially timber, usefully complemented the Egyptian economy. The Ottoman army was defeated near Konya in Anatolia (December 1832), and in 1833 the Sultan ceded the Syrian provinces to Muḥammad ʿAlī.

In 1839 Ottoman forces reentered Syria but were defeated by Ibrāhīm at the Battle of Nizip (Nezib). A fortnight later Mahmud II died, and the Ottoman Empire seemed on the verge of dissolution; it was saved only by European intervention. In 1840 Ibrāhīm was compelled to evacuate Syria. Muḥammad ʿAlī's Arabian empire (which since 1833 had extended into the Yemen) crumbled at the same time. Although in June 1841 the new sultan, Abdülmecid I (1839–61), conferred on the family of Muḥammad ʿAlī the hereditary rule of Egypt, the viceroy's powers were declining. Because of his growing senility, Ibrāhīm succeeded him (July 1848) but his reign lasted only a few months until his death the following November. The next viceroy was ʿAbbās I, the eldest grandson of Muḥammad ʿAlī. The old viceroy himself died in 1849.

Administrative changes. Muḥammad ʿAlī's military exploits would not have been possible but for radical changes in the administration of Egypt itself. Muḥammad ʿAlī was a pragmatic statesman whose principal object was to secure himself and his family in the unchallenged possession of Egypt. His immediate problem on his accession was to deal with the Mamlūks, who still dominated much of the country, and the *ʿulamāʾ*, who had helped him to power. The strength of these two groups rested largely on their control of the agricultural land of Egypt and the revenues arising therefrom. Gradually, between 1805 and 1815, Muḥammad ʿAlī eroded the system of tax farming that had diverted most of the revenues to the Mamlūks and other notables, imposed the direct levy of taxes, expropriated the landholders, and carried out a new tax survey. In 1809 he defeated the *ʿulamāʾ*, and in 1811 he massacred many of the Mamlūk leaders in Cairo, while Ibrāhīm expelled their survivors from Upper Egypt.

Restructuring of taxation

Muḥammad ʿAlī thus became effectively the sole landholder, with a monopoly over trade in crops, in Egypt, although later in his reign he made considerable grants of land to his family and dependents. The monopoly system was extended in due course from primary materials to manufactures, with the establishment of state control over the textile industry. Muḥammad ʿAlī's ambitious hopes of promoting an industrial revolution in Egypt were not realized, fundamentally because of the lack of available sources of power. The monopolies were resented by European merchants in Egypt and clashed with the economic doctrine of free trade upheld by the British government. Although a free-trade convention that was concluded between Britain and the Ottoman Empire in 1838 (the Convention of Balta Liman) was technically binding on Egypt, Muḥammad ʿAlī succeeded in evading its application up to and even after the reversal of his fortunes in 1840–41.

Reform of the military

The old-style military forces (including the Albanians), on whom Muḥammad ʿAlī relied against his internal opponents and who conquered the Hejaz, Najd, and the northern Sudan, were heterogeneous and unruly. An attempt to introduce Western methods of training in 1815 provoked a mutiny. Muḥammad ʿAlī then decided to form an army of slave-troops dependent wholly upon himself and trained by European instructors. The conquest of the Sudan was intended to provide the recruits. But the slaves, encamped at Aswān, died wholesale, and Muḥammad ʿAlī had to look elsewhere for the mass of his troops. In 1823 he took the momentous step of conscripting Egyptian peasants for the rank and file of his "new model army." On the other hand, the officers were mostly Turkish-speaking Ottomans, while the director of the whole enterprise, Sulaymān Pasha (Colonel Sève), was a former French officer. The conscription was brutally administered and military life harsh. There were several ineffective peasant revolts, while flight to the towns and (before 1831) to Syria produced rural depopulation and a decline in cultivation.

As reorganization proceeded, the viceroy gradually built a new administrative structure. While institutions were created and discarded according to his changing needs, Muḥammad ʿAlī depended essentially upon the members of his own family, particularly Ibrāhīm, and loyal servants, such as his Armenian confidant Boghos Bey. Characteristic of his governmental system were councils of officials, convened to deliberate on public business, and administrative departments (divans) that bore some resemblance to the ministries of European governments. In local administration, Muḥammad ʿAlī established a highly centralized system with a clear chain of command from Cairo through the provincial governors, down to the village headmen. Initiative was not encouraged, but firm control had taken the place of anarchy.

These changes necessitated the training of officers and officials in the new Europeanized ways of working; and this in turn resulted in the creation of a range of educational institutions alongside the traditional Muslim schools that prepared the *ʿulamāʾ*. Much of the foundation work was done by expatriates, while missions of Egyptian students were sent to Europe, especially to Paris. One of these, Rifāʿah Rāfiʿ aṭ-Ṭahṭāwī (1801–73), played the leading part in inaugurating the translation of European works into Arabic and so was a pioneer both in the interpretation of European culture to Egypt and in the renaissance of literary Arabic. The establishment of a government printing press in 1815 soon made possible the wide dissemination of the new books.

ʿAbbās I and Saʿīd, 1848–63. The reign of ʿAbbās I (1848–54) indicates how precarious was the advance of westernization in Egypt. The effort had already been relaxed in the last decade of Muḥammad ʿAlī's rule, and ʿAbbās showed himself to be a traditionalist. It was typical of his policy that he closed the school of languages and the translation bureau and sent their director, aṭ-Ṭahṭāwī, to virtual exile in the Sudan. The French, who had played so large a part in Muḥammad ʿAlī's reforms, fell into disfavour, and for diplomatic support ʿAbbās turned to their British rivals, whose support was needed against the Ottomans. Although initially ʿAbbās was ostentatiously loyal to the Sultan, he resented an attempt made at this time to curtail his autonomy. The British, for their part, had their communications with India facilitated by the

grant of a concession to build a railway from Alexandria to Cairo; the line was completed between 1851 and 1856 and was extended to Suez two years later. Saʿīd (1854–63), who succeeded on ʿAbbās' mysterious and violent death, inaugurated another reversal of policy. While he lacked Muḥammad ʿAlī's energy and ability, he was not unsympathetic to the westernizers. To his French friend Ferdinand de Lesseps (who had been a friend to Muḥammad ʿAlī as well) he granted in 1854 a concession for the cutting of a canal across the isthmus of Suez. This embroiled him both with the Sultan, whose prerogative had been encroached upon, and the British, whose overland railway route was threatened by the project; a deadlock lasted throughout his reign.

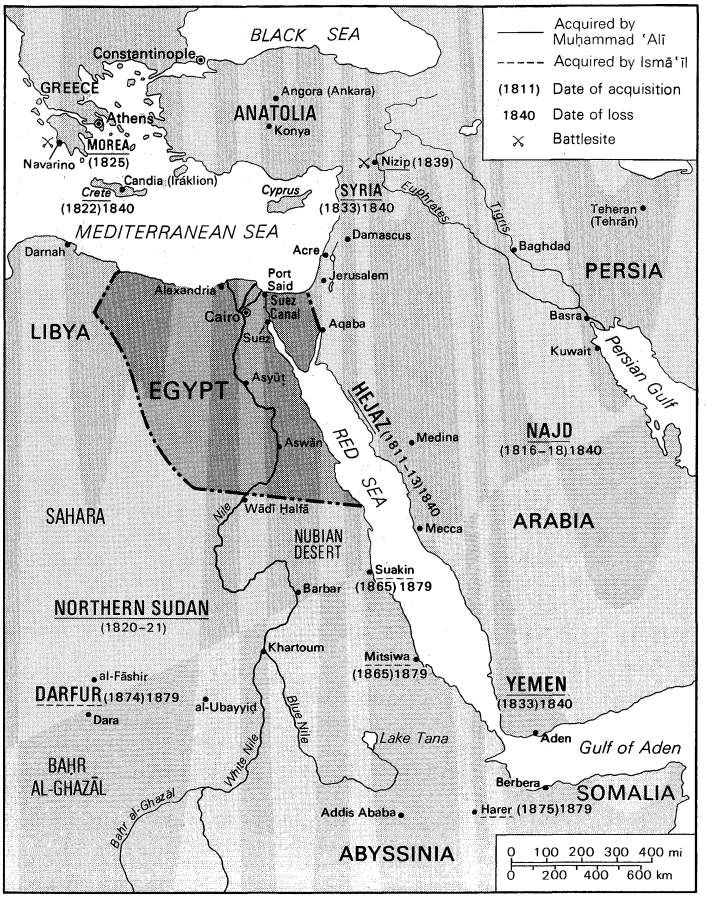
The concession to Ferdinand de Lesseps

Ismāʿīl, 1863–79. Ismāʿīl, the son of Ibrāhīm Pasha, who succeeded on the death of Saʿīd, displayed some of his grandfather's dynamic energy and enthusiasm for modernization. He lacked caution, however, and his reign ended in catastrophe. From his predecessors he inherited a precarious economy and a burden of debt. The American Civil War (1861–65) produced a boom in Egyptian long-staple cotton. This had been introduced and developed in Muḥammad ʿAlī's time, but its production had languished until the interruption of supplies of American cotton caused a fourfold increase in price during the war years. When peace returned, prices collapsed with disastrous consequences for the Egyptian economy. In the management of his finances, Ismāʿīl was both extravagant and unwise and laid himself open to unscrupulous exploitation. Ismāʿīl was committed to the Suez Canal project, but he modified the grant in two important respects: by withdrawing the cession of a strip of land from the Nile to the Suez isthmus, along which a freshwater canal was to be constructed, and by refusing to provide unlimited peasant labour for the project. The matter was submitted to arbitration; an indemnity of more than £3,000,000 was imposed on Ismāʿīl, who also agreed to pay for a large block of shares put by de Lesseps to Saʿīd's account. French pressure on the Sultan succeeded at last in overcoming resistance to the canal project at Istanbul, and a firman (decree from the sultan) authorizing its construction was granted in March 1866. Work had in fact already been going on for seven years, and in November 1869 the Suez Canal was opened to shipping by the empress Eugénie, the wife of Napoleon III of France. The incident symbolized the political and cultural orientation of Egypt in the middle decades of the 19th century.

Ismāʿīl, in other ways, presented himself as the ruler of a new and important state. Although his relations with his suzerain, Sultan Abdülaziz (1861–76), were normally friendly, he was no less anxious than his predecessors to secure the autonomy of his dynasty. In 1866 he obtained a firman establishing the succession by primogeniture in his own line—abandoning the contemporary Ottoman rule of succession by the eldest male. A year later a firman conferred upon Ismāʿīl the special title of khedive, which had in fact been used unofficially since Muḥammad ʿAlī's time and which distinguished the viceroy of Egypt from other Ottoman governors. A period of strained relations developed between the Khedive and the Sultan arising from Ismāʿīl's implied pretensions to sovereignty at the time of the opening of the Suez Canal in 1869, but the two were later reconciled; a firman reconfirmed the Khedive's privileges in 1873. These concessions by the Sultan, however, cost Ismāʿīl heavy expenditure and an increase in the annual Egyptian tribute and formed another factor in the growth of Ismāʿīl's indebtedness.

Ismāʿīl had inherited an African empire in the northern Sudan. Since the middle of the century, in consequence of the abolition of the monopolies, merchants had penetrated south and southwest, up the White Nile and the Baḥr al-Ghazāl, in search of ivory. An ancillary slave trade had developed that was repugnant to the European conscience. Humanitarian and expansionist motives thus coincided to persuade Ismāʿīl to extend Egyptian rule into these remoter regions. He made considerable use of expatriates, notably the Englishmen Sir Samuel Baker and Charles George ("Chinese") Gordon, who extended the Khedive's nominal authority to the African Great Lakes.

Penetration southward



Expansion of Egypt under Muḥammad ʿAlī and Ismāʿīl.

Another series of events led to the conquest in 1874 of the sultanate of Darfur in the west. The Khedive also wished to make Egypt the dominant power in the Red Sea region. The Sultan granted him the old Ottoman ports of Suakin and Mitsiwa in 1865. Egyptian control was established on the Somali coast, and in 1875 Harer was captured. Attempts to invade Abyssinia in 1875 and 1876 were, however, unsuccessful and marked the limits of Ismāʿīl's imperial expansion.

Like other parts of the Ottoman Empire, Egypt was bound by the Capitulations—a system of privileges derived from ancient treaties with former sultans. Under the Capitulations, European and American residents in Egypt were exempt from local taxation and were subject only to their own consular courts. By patient negotiations over several years, Nūbār Pasa, Ismāʿīl's Armenian minister, succeeded in establishing the Mixed Courts in 1875. These had jurisdiction in cases involving Egyptians and foreigners, or foreigners of different nationalities, and had both foreign and Egyptian judges, who administered codes based on French law.

By this time the social consequences of the agrarian and political changes inaugurated by Muḥammad ʿAlī were clearly appearing. The Khedive and his family were the principal landholders in Egypt, possessing extensive personal estates quite apart from the state lands. Around the khedivial family was a parvenu aristocracy that held the principal civil and military offices. Many of its members were also great landowners; most of them were Turkish or Circassian by origin. Although the condition of the peasantry had been adversely affected by military conscription, by corvées for public works (including large-scale demands for labour on the railways and the Suez Canal), and by ill-considered economic and industrial experiments, the rights of cultivators on their land gradually increased. The richer peasants, from whom the village headmen were recruited, in particular increased in importance. When in November 1866 Ismāʿīl set up the consultative council known as the Assembly of Delegates, the members of which were chosen

Creation of the Assembly of Delegates

by indirect election, the great majority of those chosen were village headmen. While Ismāʿīl did not intend that the Assembly should limit his power, its establishment and composition were indications of the political development of the Egyptians in 60 years. Conscription had affected the political significance of the army. The ascendancy of the entrenched Turco-Circassians was challenged by native Egyptian officers, who resented the privileged position of their foreign colleagues. The defeat of the Circassian commander in chief, Rātib Pasha, by the Abyssinians in 1876 was a blow from which the prestige of the old officer group never recovered.

In the Assembly and the army, and among the westernized intelligentsia, politically conscious individuals and groups began to emerge who drew their ideas from both Western and Islāmic sources. Their organization was for the most part small-scale and ephemeral, and their outlook was subversive, being hostile to the autocracy of the Khedive, the ascendancy of the Turco-Circassians, and the pervasive power of the Europeans.

Political tension increased in the last years of Ismāʿīl's reign. Various expedients to postpone bankruptcy (*e.g.*, the sale in 1875 of his Suez Canal shares to Britain) had failed, and in 1876 the Caisse de la Dette Publique (Commission of the Public Debt) was established for the service of the Egyptian debt. Its members were nominated by France, Britain, Austria, and Italy. In the same year, Egyptian revenue and expenditure were placed under the supervision of a British and a French controller (the Dual Control). After an international enquiry in 1878, Ismāʿīl accepted the principle of ministerial responsibility for government and authorized the formation of an international ministry under Nūbār. Ismāʿīl, however, was not prepared to yield his autocracy tamely. In 1879 he profited from an army demonstration against the European ministers to dismiss Nūbār, and he worked in alliance with the Assembly of Delegates to destroy international control over Egypt. By this time, however, his standing outside Eygpt had been lost; and in June 1879, Sultan Abdülhamid II (1876–1908), at the instigation of France and Britain, deposed him in favour of his son, Tawfīq.

Renewed European intervention, 1879–82. European domination was immediately reasserted. The Dual Control was revived, the British controller being Evelyn Baring. By the Law of Liquidation (July 1880), the annual revenues were divided into two approximately equal portions, one of which was assigned to the Caisse de la Dette. The Assembly of Delegates was dissolved. The forces of resistance that Ismāʿīl had stimulated were not, however, allayed by these means. There had already come into existence a nationalist group within the Assembly, prominent among whom was Sharīf (Cherif) Pasha, prime minister from April to August 1879. In the army a group of Egyptian officers, whose leader was ʿUrābī (Arabi) Pasha, was disaffected from the Khedive and resentful of European control of Egypt. By 1881 these two groups had allied to form the National Party, al-Ḥizb al-Waṭanī.

Open tension appeared with a petition drawn up in January 1881 by ʿUrābī and two of his colleagues against the war minister, Rifqī Pasha, a Circassian. They were arrested and court-martialed but released by mutineers. Tawfīq capitulated, dismissed Rifqī, and appointed Bārūdī Pasha, one of ʿUrābī's friends, as war minister. But the ʿUrābists still felt themselves endangered; a military demonstration in Cairo in September 1881 compelled Tawfīq to appoint a new ministry under Sharīf and to convoke the Assembly. But the alliance between the military group and Sharīf was uneasy.

European intervention

Meanwhile, the European powers were becoming increasingly alarmed. A joint English and French communication sent in January 1882 with the intention of strengthening the Khedive against his opponents had the contrary effect. The Assembly of Delegates swung toward the ʿUrābists. Sharīf resigned and Bārūdī became prime minister with ʿUrābī as war minister. Rioting ensued on June 11 after British and French naval forces had been sent to Alexandria. From this point Britain took the initiative. The French refused participation in a bombardment of Alexandria (July 11), while an international conference held at Istanbul was boycotted by the Turks and produced no solution of the problem. The British government finally resolved on intervention and sent an expeditionary force to the Suez Canal. The ʿUrābists were rapidly defeated at Tall al-Kabīr (Sept. 13, 1882), and Cairo was occupied the next day.

THE PERIOD OF BRITISH DOMINATION (1882–1952)

The British occupation and the Protectorate (1882–1922). The British occupation marked the culmination of developments that had been at work since 1798: the de facto separation of Egypt from the Ottoman Empire, the attempt of European powers to influence or control the country, and the rivalry of France and Britain for ascendancy in the country. Through the last minute withdrawal of the French, the British had secured the sole domination of Egypt. W.E. Gladstone's Liberal government was, however, reluctant to prolong the occupation or to establish formal political control, which it feared would antagonize both the Sultan and the other European powers; but the British were unwilling to evacuate Egypt without securing their strategic interests, and this never seemed possible without maintaining a military presence there.

An incident at the outset of the occupation was significant of future tensions. On British insistence, the Khedive's government was obliged to place ʿUrābī and his associates on public trial and to commute the resulting death sentences to exile. Tawfīq's prestige, slight enough at his accession, and diminished in the three years before the occupation, was still further undermined by this intervention of the British government. Meanwhile, Lord Dufferin, the British ambassador in Istanbul, visited Egypt and prepared a report on measures to be taken for the reconstruction of the administrative system. The projects of reform that he envisaged would necessitate an indefinite continuation of the occupation. The implications of this for British policy were slowly and reluctantly accepted by the ministry in London, under pressure from its representative in Cairo, the British agent and consul general, Sir Evelyn Baring, who in 1891 became Lord Cromer.

Public trial of ʿUrābī

Two principal problems confronted the occupying power: first, the acquisition of some degree of international recognition for its special but ambiguous position in Egypt; second, a definition of its relationship to the khedivial government, which formed the official administration of the country. The main European opponents of recognition of the British position were the French, who resented the abolition of the Dual Control (December 1882). The Caisse de la Dette remained in existence, and until 1904 the British had to tread warily in order to circumvent French opposition in this institution. In the early years of the occupation, when Egyptian finances were in disarray, French hostility was a serious problem, but from 1889 onward there was a budget surplus and consequently greater freedom of action for the Egyptian government. A moderate degree of international agreement over Egypt was attained by the Convention of London (1885), which secured an international loan for the Egyptian government and added two further members (nominated by Germany and Russia) to the Caisse de la Dette. In 1888 the Convention of Constantinople (Istanbul) provided that the Suez Canal should always be open in war and peace alike. This was, however, a statement of principle rather than fact; without British cooperation it remained a dead letter.

In matters concerning the international status of Egypt, the decisions were taken in London, but where the internal administration of the country was concerned, Cromer's opinions were usually conclusive. Although throughout the occupation the facade of khedivial government was retained, British advisers attached to the various ministries were more influential than their ministers, while Cromer himself steadily increased his control over the whole administrative machine. (P.M.Ho.)

Cromer's personal rule

Tawfīq himself gave little trouble, but his prime ministers were more tenacious. Sharīf Pasha, prime minister at the beginning of the occupation until 1884, and his successors Nubar Pasha (1884–88) and Riyāḍ (Riaz) Pasha (1888–91) resigned because of clashes over administrative control. Thereafter, until November 1908, with a break in 1893–

95, the prime minister was Muṣṭafā Fahmī Pasha, who showed himself an obedient instrument of Cromer.

ʿAbbās Ḥilmī II, 1892–1914. The death of Tawfīq and the accession of his 17-year-old son, ʿAbbās Ḥilmī II, in 1892 marked the beginning of a new phase of opposition to the occupation. The new khedive was not content to accept Cromer's tutelage, while the British agent resented the attempts of one so much his junior to play a serious role in Egyptian politics. ʿAbbās dismissed Muṣṭafā Fahmī in January 1893 and tried to appoint his own nominee as prime minister. Cromer, backed by the British government, frustrated his endeavours, and Fahmī returned to office in November 1895. ʿAbbās provoked another crisis in January 1894 by public criticism of British military officers and especially H.H. Kitchener, the sirdar (commander in chief). Once again Cromer intervened, and ʿAbbās was compelled to make amends.

Other considerations apart, the behaviour of ʿAbbās in the early years of his reign indicated the emergence of a new generation who had only been children when the occupation began. One of ʿAbbās' contemporaries was Muṣṭafā Kāmil (1874–1908), who had studied in France and then had entered a circle of Anglophobe writers and politicians. On returning to Egypt in 1894 he had reached an understanding with the Khedive on the basis of their common detestation of the British occupation. By his speeches and writings (in 1900 he founded his own newspaper, *al-Liwā*), he endeavoured to create an Egyptian patriotism that would rally the entire nation around the Khedive. A boost was given to nationalism by the campaigns for the reconquest of the Sudan (1896–98) and still more by the Condominium Agreement of 1899, which nominally gave Egypt and Britain joint responsibility for the administration of the reconquered territory but in effect made the Sudan a British possession.

Muṣṭafā Kāmil

A final episode in the reconquest of the Sudan, the confrontation of British and French at Fashoda on the White Nile in 1898, was followed by the reconciliation of the two powers in the Entente Cordiale (1904), which in effect gave Britain a free hand in Egypt. This was a blow to the hopes of Muṣṭafā Kāmil and to his alliance with the Khedive, who showed himself more willing to cooperate with Cromer. Muṣṭafā Kāmil now turned to Sultan Abdülhamid. When a dispute (the Ṭābah Incident, 1906) arose between the Ottomans and the occupying power over the Sinai Peninsula, Muṣṭafā Kāmil sought to rally Egyptian nationalist opinion in favour of the Sultan, but Muṣṭafā Kāmil died in 1908.

British domination in Egypt and Cromer's personal ascendancy never seemed more secure than in the period following the Entente Cordiale. But the "veiled protectorate" had hidden weaknesses. Cromer was both out of touch and out of sympathy with the new generation of Egyptians. The occupation had become to all intents and purposes permanent, and the consequent growth of the British official establishment created frustration among educated Egyptians. The British, however, saw themselves as the benefactors of the Egyptian peasantry, whom they had delivered from the corvée and the lash. The Dinshawāy Incident showed them in another light. In June 1906 a fracas between villagers at Dinshawāy and a party of British officers out pigeon shooting resulted in the death of a British officer. The special tribunal set up to try the matter imposed exemplary and brutal sentences on the villagers. In the bitter aftermath of this affair, Cromer retired in May 1907.

The Dinshawāy Incident

Sir Eldon Gorst, who succeeded Cromer, had served in Egypt from 1886 to 1904 and brought a fresh mind to bear on the problems of the occupation. He obtained an understanding with the Khedive and endeavoured to diminish the growing power and numbers of the British establishment. At the same time he tried to give more effective authority to Egyptian political institutions. Muṣṭafā Fahmī's long premiership ended and he was followed by a Copt, Buṭrus Ghālī Pasha. When Gorst died prematurely in July 1911, he had attained only limited success. Many British officials resented his policies, which at the same time failed to conciliate the nationalists. A project for the extension of the Suez Canal Company's 99-year concession by 40 years was thrown out by the General Assembly (a quasi-parliamentary body, set up in 1883), while Buṭrus Ghālī, who had advocated it, was assassinated a few days later by a Muslim extremist. The appointment of Lord Kitchener to succeed Gorst portended the end of conciliation of the Khedive. But Kitchener, although autocratic, was not wholly conservative; his attempts to limit the power and influence of ʿAbbās Ḥilmī served the interests of the nationalists. The Organic Law of 1913 created a new and more powerful Legislative Assembly that served as a training ground for the nationalist leaders of the postwar period. At the same time, the peasants were helped by improved agriculture and by legal protection of their holdings from seizure for debt.

World War I and independence. In November 1914 Britain declared war on the Ottoman Empire and in December proclaimed a protectorate over Egypt, deposed ʿAbbās, and appointed his uncle, Ḥusayn Kāmil, with the title of sultan. Kitchener was succeeded by Sir Henry MacMahon, and he by Sir Reginald Wingate, both with the title of high commissioner. Although Egypt was not required to provide troops, the people, and particularly the peasantry, suffered from the effects of war. The declaration of martial law and the suspension of the Legislative Assembly curbed the activities of middle-class nationalists. Ḥusayn Kāmil died in October 1917 and was succeeded by his ambitious brother, Aḥmad Fu'ād.

On Nov. 13, 1918, two days after the Armistice, Wingate was visited by three Egyptian politicians headed by Sa'd Zaghlūl Pasha. Zaghlūl demanded autonomy for Egypt and announced his intention of leading a delegation (Arabic *wafd*) to state his case in England. The British government's refusal to accept a delegation, followed by the arrest of Zaghlūl, produced a widespread revolt in Egypt; and Lord Allenby, the victor over the Turks in Palestine, was sent out as special high commissioner. Allenby insisted on concessions to the nationalists in the hopes of reaching a settlement. Zaghlūl was released, and the Wafd, now a countrywide organization, dominated Egyptian politics. The Milner Commission (1919–20), sent to report on the establishment of constitutional government under the protectorate, was boycotted, but Milner subsequently had private talks with Zaghlūl in London. Finally, hoping to outmaneuver Zaghlūl and to build up a group of pro-British politicians in Egypt, Allenby pressed his government to promise independence without previously securing British interests by a treaty. The declaration of independence (Feb. 28, 1922) ended the protectorate but, pending negotiations, reserved four matters to the discretion of the British government: the security of imperial communications, defense, the protection of foreign interests and of minorities, and the Sudan. On March 15 the Sultan became King Fu'ād I of Egypt.

Attempt of Zaghlūl to secure Egypt's independence

The Kingdom of Egypt (1922–52). The new kingdom was in form a constitutional monarchy. The constitution, based on that of Belgium and promulgated in April 1923, defined the King's executive powers and established a bicameral legislature. An electoral law provided for universal male suffrage and the indirect election of deputies to the lower house: the Senate was half elected and half appointed. But Egyptian constitutionalism was as illusory as Egyptian independence. A political struggle was continually waged among three opportunist contestants—the King, the Wafd, and the British.

The interwar period. Fu'ād was never popular and felt insecure, and was therefore prepared to intrigue with the nationalists or with the British to secure his position and powers. The Wafd, with its mass following, elaborate organization, and (until his death in 1927) charismatic leader in Zaghlūl, was the only truly national party in Egypt. Ideologically, it stood for national independence against British domination and for constitutional government against royal autocracy. In practice—and increasingly as time went on—its leaders were prepared to make deals with the British or the King to obtain or retain power. Personal and political rivalries led to the formation of splinter parties, the first of which, the Liberal Constitutionalist Party, broke off as early as 1922. The primary aim of the British government, represented by its high

commissioner (after 1936, its ambassador), was to secure imperial interests, especially the control of communications through the Suez Canal. The need for a treaty to safeguard these interests led Britain on more than one occasion to conciliate nationalist feeling by supporting the Wafd against the King.

Political crises of the 1920s

The first general election, in January 1924, gave the Wafd a majority, and Zaghlūl became prime minister for a few months marked by unsuccessful treaty discussions with the British and tension with the King. When in November 1924 Sir Lee Stack, the sirdar and governor-general of the Sudan, was assassinated in Cairo, Allenby immediately presented an ultimatum that, though later modified by the British government, caused Zaghlūl to resign. The general election of March 1925 left the Wafd still the strongest party, but the Parliament no sooner met than it was dissolved. For more than a year Egypt was governed by decree. The third general election, in May 1926, again gave the Wafd a majority. The British frowned on a return of Zaghlūl to the premiership, and the office went instead to the Liberal Constitutionalist ʿAdlī Yegen (Yakan), while Zaghlūl held the presidency of the Chamber of Deputies until his death in 1927. Once again tension developed between the Parliament and the King, and in April 1927 ʿAdlī resigned, to be succeeded by another Liberal Constitutionalist, ʿAbd al-Khāliq Tharwat (Sarwat) Pasha, who negotiated a draft treaty with the British foreign secretary. The draft treaty, however, failed to win the approval of the Wafd. Tharwat resigned (March 1928), and Muṣṭafā an-Naḥḥās (Nahas) Pasha, Zaghlūl's successor, became prime minister. But the King dismissed him in June and dissolved the Parliament in July. In effect, the constitution was suspended, and Egypt was again governed by decree under a Liberal Constitutionalist premier, Muḥammad Maḥmūd Pasha.

Draft treaty proposals were agreed upon in June 1929, but since Maḥmūd was unable to overcome Wafdist opposition, British influence was thrown behind a return to constitutional government, hoping that a freely elected Parliament would approve the proposals. In the fourth general election (December 1929), the Wafd won a majority, and an-Naḥḥās again became prime minister. Resumed treaty negotiations broke down over the problem of the Sudan, from which the Egyptians had been virtually excluded since 1924. An-Naḥḥās also clashed with the King, whose influence he sought to curtail. He resigned in June 1930, and Fu'ād appointed Ismāʿīl Ṣidqī (Sidki) Pasha to the premiership. The constitution of 1923 was abrogated, and another was promulgated by royal decree. This, with its accompanying electoral law, strengthened the King's power. By this and other measures, Ṣidqī sought to break the power of the Wafd, which boycotted the general election of June 1931. The strong government of Ṣidqī lasted until September 1933, when he was dismissed by the King. Thereafter, for more than a year, palace-appointed governments ruled Egypt.

But Fu'ād, whose health was failing, could not hold out indefinitely against the internal pressure of the Wafd and the external pressure of Britain, which was becoming increasingly anxious for a treaty with Egypt. In April 1935 the constitution of 1923 was restored, and a general election in May 1936 gave the Wafd a majority once more. Fu'ād had died in the previous month and was succeeded by his son Farouk (Fārūq), still a minor. An-Naḥḥās became prime minister for the third time. Agreement was quickly reached with Britain, and a treaty of mutual defense and alliance was signed in August 1936. At the conference of Montreux, held in the following year, Egypt, with the backing of Britain, obtained the immediate abolition of the Capitulations and the extinction of the Mixed Courts after 12 years. In 1937 also, Egypt became a member of the League of Nations.

Treaty with Britain

An-Naḥḥās had reached the height of his power, but he was soon to be overthrown. In July 1937 the young king Farouk came of age and assumed his full royal powers. He was both popular and ambitious to rule, and tension rapidly developed between him and his prime minister. A split developed in the Wafd: Maḥmūd Fahmī an-Nuqrāshī (Nokrashy) Pasha and Aḥmad Māhir (Maher) Pasha were expelled and formed the Saʿdist Group. The Wafdist youth movement, known as the Blueshirts, was opposed by the Greenshirts of Young Egypt, a fascist organization. In December 1937 King Farouk dismissed an-Naḥḥās. In the ensuing general election (April 1938), the Wafd won only 12 seats.

World War II and its aftermath. Although at the outbreak of World War II in September 1939 Egypt provided facilities for the British war effort, few Egyptians were enthusiastic supporters of Britain and many expected its defeat. In 1940 the British brought pressure on the King to dismiss his prime minister, ʿAlī Māhir, and to appoint a more cooperative government. When, early in 1942, German forces prepared to invade Egypt, a second British intervention compelled King Farouk to accept an-Naḥḥās as prime minister. The Wafd, its power confirmed by overwhelming success in the general election of 1942, cooperated with Britain. Nevertheless, the intervention of February 1942 had disastrous consequences. It confirmed Farouk's hostility to both the British and an-Naḥḥās and tarnished the Wafd's pretensions as the standard-bearer of Egyptian nationalism. The Wafd was damaged also by internal rivalries.

An-Naḥḥās was dismissed by the King in October 1944. His successor, Aḥmad Māhir, was acceptable to the British, but he was assassinated in February 1945, at the moment of Egypt's declaration of war on Germany and Japan. He was succeeded by a Saʿdist, an-Nuqrāshī Pasha.

At the end of World War II, Egypt was in a thoroughly unstable condition. The Wafd declined and its political opponents took up the nationalist demand for a revision of the treaty of 1936—in particular for the complete evacuation of British troops from Egypt and the ending of British control in the Sudan. Politics was passing into the hands of radicals. The Muslim Brotherhood, founded in 1928, developed from an orthodox Islāmic reformist movement into a militant mass organization. Demonstrations in Cairo became increasingly frequent and violent. The pressure rendered it impossible for any Egyptian government to attempt a settlement of its two main external problems: the need to revise the treaty with Britain, and the wish to support the Arab cause in Palestine. Negotiations with Britain, undertaken by an-Nuqrāshī and (after February 1946) by his successor, Ṣidqī, broke down over the British refusal to prejudice the possible independence of the Sudan. Although Egypt referred the dispute to the United Nations in July 1947, the deadlock continued.

Anti-British demonstrations

Until the interwar period neither the Egyptian public nor the politicians had shown much interest in Arab affairs generally; Egyptian nationalism had developed as an indigenous response to local conditions. After 1936, however, Egypt became involved in the Palestine problem, and in 1943–44 it played a leading part in the formation of the Arab League. After World War II, Egypt became increasingly committed to the Arab cause in Palestine, but its unexpected and crushing defeat in the first Arab–Israeli War (1948–49), which had been launched with Syria, Iraq, and Jordan in response to the declaration of the State of Israel in May 1948, contributed to disillusionment and political instability. The Muslim Brotherhood increased its terrorist activities. An-Nuqrāshī, again prime minister, endeavoured to suppress the organization and was assassinated in December 1948. The Brotherhood's leader was murdered two months later.

A general election in January 1950 gave the Wafd a majority, and an-Naḥḥās again formed a government. Failing to reach agreement with Britain, in October 1951 he abrogated both the 1936 treaty and the Condominium Agreement of 1899. Anti-British demonstrations were followed by guerrilla warfare against the British garrison in the Canal Zone. British military action in Ismailia was followed on Jan. 26, 1952, by the burning of Cairo by demonstrators. An-Naḥḥās was dismissed, and there were four prime ministers in the ensuing six months.

(P.M.Ho./Ed.)

THE REVOLUTION AND THE REPUBLIC

The Nasser regime. At mid-century Egypt was ripe for revolution. Political groupings of both right and left pressed

for radical alternatives. From an array of contenders for power, it was a movement of military conspirators—the Free Officers led by Col. Gamal Abdel Nasser—that toppled the monarchy in a coup in July 1952. In broad outline, the history of contemporary Egypt is the story of this coup, which preempted a revolution but then itself became a revolution from above. For three decades rule by Free Officers brought just enough advance at home and enhancement of standing abroad to make Egypt an island of stability in a turbulent Middle East.

The Free Officers

The coup of July 1952 was fueled by a powerful but vague Egyptian nationalism rather than by a coherent ideology. It yielded a regime whose initially reformist character was given more precise form by a domestic power struggle and by the necessity of coming to terms with the British, who still occupied their base at Suez.

The domestic challenge to Nasser came in February–April 1954 from Gen. Mohammad Naguib, an older officer who served as figurehead for the Free Officers. Political parties had been abolished in January 1953. To supplement his power base in the military forces, Nasser drew on the police and on working-class support mobilized by the newly created mass political organization called the National Union. The small middle class, the former political parties, and the Muslim Brotherhood all rallied to Naguib. Nasser's triumph meant that a strong reliance on the military and security apparatus, coupled with carefully controlled manipulation of the civilian population, would be basic to the new system of rule.

Nasser's initial moderation

Obscured in the West was Nasser's initial moderation regarding Egypt's key foreign policy challenges—the Sudan, the British presence, and Israel. An agreement signed in 1954 established a transitional period of self-government for the Sudan, which became an independent republic in January 1956. Prolonged negotiations yielded the Anglo-Egyptian Treaty of 1954, under which British troops were to be evacuated gradually from the Canal Zone. Some Egyptians were critical, finding the treaty unsatisfactory from an Egyptian nationalist perspective. An attempt to assassinate Nasser by a member of the Muslim Brotherhood in October 1954 was used as a pretext to crush that organization.

In retrospect, it is clear that Nasser was the reluctant champion of the Arab struggle against Israel. Domestic development was his priority. A dangerous pattern of violent interactions, however, was evolving that would eventually draw the Egyptians into conflict with Israel. Small groups of Palestinian raiders, including some operating from Egyptian-controlled Gaza, were infiltrating Israel's borders. In October 1953 the Israeli government initiated the policy of large-scale retaliation that it pursued thereafter. One such strike—an attack on Gaza in February 1955 that left 38 Egyptians dead—exposed the military weakness of the Free Officer regime.

In September 1955 Nasser announced that an arms agreement had been signed between Egypt and Czechoslovakia (acting for the Soviet Union). The way to improved Soviet–Egyptian relations had been prepared by Nasser's refusal to join the Baghdad Pact (the Middle East Treaty Organization, later known as the Central Treaty Organization), which had been formed earlier that year by Turkey, Iraq, Iran, Pakistan, and the United Kingdom, with the support of the United States, to counter the threat of Soviet expansion. With the arms agreement of 1955, the Soviet Union eluded efforts to contain its actions and established itself as a force in the Middle East.

The erosion of Nasser's initially pro-Western orientation was accelerated further by the denial of funds previously promised by the United States and Britain for the construction of a high dam at Aswān. Defiantly, Nasser announced the nationalization of the Suez Canal Company in 1956 to finance the dam. In its subsequent attack on Egypt in October 1956, Israel was joined by the British, who were enraged by the nationalization, and the French, who were angered by Egyptian aid to the revolt in Algeria. Pressure on the invading powers by the United States and the Soviet Union, however, soon ended the so-called Suez War, leaving Nasser triumphant (despite his military losses) and with the Suez Canal firmly in Egyptian hands.

The Suez War of 1956

Nasser, who had been elected president in June 1956, pursued a more radical line in the decade following the Suez War. He launched an ambitious program of domestic transformation, a revolution from above that was paralleled by a drive for Egyptian leadership in the Arab world. Early in 1958 Egypt combined with Syria to form the United Arab Republic (U.A.R.), but it was a reluctant marriage of convenience and was dissolved in bitterness in September 1961 (Egypt retained the name United Arab Republic until 1971). The secession of Syria was blamed by Nasser on Syrian "reactionaries," and in direct response he pushed the revolution in Egypt further to the left. The following spring a National Charter proclaimed that Egypt's would be a regime of "scientific socialism" with a new mass organization, the Arab Socialist Union (ASU), to function in place of the National Union.

Impressive domestic gains were registered. In 1950 industry contributed 10 percent to the total national output; by 1970 that figure had increased to 21 percent. Unfortunately, these achievements in industry were not matched in agriculture, and they were further undercut by rapid population growth.

Throughout this period the potential military danger from Israel was a constant factor in the calculations of the U.A.R. government. It was a motive in strengthening ties with the Soviet bloc and producing a series of initiatives for cooperation among the Arab states, which, however, were disappointing. Nasser masked essential Egyptian moderation on the Israeli issue with a militant rhetoric of confrontation that was necessary to preserve his standing in the Arab world.

The failure of the union with Syria had been a blow to Nasser's pan-Arab standing. To regain the initiative, Nasser intervened in 1962–67 on the republican side of the Yemeni civil war. That intervention provoked conflict with Saudi Arabia, which supported the Yemeni royalists, and with the United States, which in turn supported the Saudis. Until then, Nasser had managed to obtain impressive aid from both the Soviet Union and the United States. Because of Egyptian intervention in the Yemen, U.S. aid was cut off in the mid-1960s.

Egyptian intervention in Yemen

This series of reversals was one key factor in the mood of desperation that pushed Nasser to abandon his policy of "militant inaction" toward Israel. For 10 years relative peace on the border with Israel was precariously maintained by the presence of a UN Emergency Force (UNEF) stationed on the Egyptian side. In the Arab summit conferences of the early 1960s Nasser had counseled restraint, but in 1966 events eluded his control. Palestinian incursions against Israel were launched with greater frequency and intensity from bases in Jordan, Lebanon, and, especially, Syria. A radical Syrian regime openly pledged support to the Palestinian guerrilla raids. On Nov. 13, 1966, an Israeli strike into Jordan left 18 dead and 54 wounded. Taunted openly for hiding behind the UNEF, Nasser was forced to act. The Egyptian president requested the withdrawal of the UNEF from the Sinai border. But that would include, as the United Nations interpreted the order, the removal of UN troops stationed at Sharm ash-Shaykh at the head of the Gulf of Aqaba. The posting of Egyptian troops there would mean the closing of the gulf to the Israelis.

Israel had made it clear that the closing of the gulf would be a cause for war. On June 5, 1967, Israel launched a preemptive attack on Egypt and Jordan later known as the June (or Six-Day) War. All of Egypt's airfields were struck, and the bulk of Egyptian planes were demolished on the ground. In the Sinai, Egyptian forces were defeated and put to flight. An estimated 10,000 Egyptians died. The Israelis reached the Suez Canal on June 9. Egypt was crushed and Nasser resigned. A popular outpouring of support, only partially manipulated by the government, refused the President's resignation. But the Nasser era was, in fact, over. In both domestic and foreign affairs, Nasser began a turn to the right that his successor, Anwar el-Sādāt, was to accelerate sharply.

Six-Day War of 1967

The Sādāt regime. Nasser died on Sept. 28, 1970, and was succeeded by his vice president, Sādāt, himself a Free Officer. Although regarded at the time as an interim figure,

Sādāt soon revealed unexpected gifts for political survival. In May 1971 he outmaneuvered a formidable combination of rivals for power, calling his victory the "Corrective Revolution." Sādāt then used his strengthened position to manage a war with Israel in October 1973, thereby setting the stage for a new era in Egypt's history.

October War of 1973

The Sādāt era really began with the October (or Yom Kippur) War of 1973. The concerted Syrian–Egyptian surprise attack on October 6 surprised not only Israel but also the world. There were no illusions that Israel could be vanquished. Rather, the war was launched with the diplomatic aim of convincing a chastened, if still undefeated, Israel to negotiate on terms more favourable to the Arabs. Preparation for the war involved a loosening of ties with the Soviet Union; to that end, in July 1972 Sādāt had announced the withdrawal of all Soviet military advisers who, it was claimed later, had opposed the Egyptian determination to fight.

Egypt did not win the war of 1973 in any conventional sense. As soon as Israel recovered from the initial shock of Arab gains in the first few days of fighting—and once the United States abandoned its early equivocation and resupplied Israel with a massive airlift—the Israelis demonstrated their military superiority. A cease-fire was secured by the United States.

Still, the initial successes in October 1973 were sufficient to allow Sādāt to pronounce the war an Egyptian victory and to openly and honourably seek peace. Egyptian interests, as Sādāt saw them, dictated peace with Israel. Despite almost immediate difficulties with his Syrian allies, Sādāt signed the Sinai I (1974) and Sinai II (1975) disengagement agreements that returned Sinai and secured large foreign assistance commitments to Egypt. When Israeli inflexibility combined with Arab resistance to slow events, Sādāt made his dramatic journey to Jerusalem on Nov. 19, 1977, to address the Israeli Knesset (Parliament). The subsequent meeting in September 1978 of Sādāt, Israeli Prime Minister Menachem Begin, and U.S. Pres. Jimmy Carter at Camp David, Md., led directly to the Israeli–Egyptian peace treaty of March 26, 1979. The treaty provided for Egyptian–Israeli normalization and established a framework for the Palestinian issue. Its provisions included the withdrawal of Israeli armed forces and civilians from Sinai within three years, special security arrangements in Sinai, a buffer zone along the Sinai–Israel border to be manned by United Nations peacekeeping forces, the exchange of ambassadors, and the establishment of normal economic and cultural relations. The status of the Israeli-occupied West Bank and Gaza territories and the question of Palestinian autonomy were to be negotiated.

Peace with Israel

Sādāt linked his peace initiative to the task of economic reconstruction, and an open-door policy was proclaimed. A liberalized Egyptian economy would be revitalized by the inflow of Western and Arab capital. The peace process did produce economic benefits, notably a vast U.S. aid program, begun in 1975, that reached more than $1,000,000,000 a year by the 1980s.

The Sādāt peace with Israel was not without its costs, however. As the narrowness of the Israeli interpretation of Palestinian autonomy under the Camp David agreement became clear, Sādāt found it impossible to convince the Arab world that the accords dealt justly with legitimate Palestinian rights. Egypt lost the financial support of the Arab states and, shortly after signing the peace treaty, was expelled from the Arab League.

At home democratization of political life did not prove to be an acceptable substitute for economic revitalization. On Jan. 18–19, 1977, demonstrations provoked by economic hardship broke out in Egypt's major cities. An estimated 79 persons were killed, 1,000 were wounded, and 1,250 were jailed. The removal of the most oppressive features of Nasser's rule, the return in controlled form to a multiparty system, and (at least initially) the Sādāt peace with Israel were all welcomed. But, as Egypt entered the 1980s, the lack of progress on the Palestinian issue and the failure to relieve mass economic hardships, heightened by widening class gaps, threatened to destabilize the Sādāt regime. In the West, Sādāt's international role initially obscured these danger signs. Then, in September 1981, his arrest of more than 1,300 of the political elite of Egypt signaled the precariousness of his position.

Egypt after Sādāt. Sādāt's assassination on Oct. 6, 1981, by members of the radical fringe of the Muslim religious opposition was greeted in Egypt by a deafening calm. It was with a profound sense of relief that Egyptians brought Hosni Mubarak, Sādāt's handpicked vice president, to power with a mandate for cautious change. As an air force general and hero of the October War, Mubarak had played an important role in Sādāt's rule from the early 1970s. (R.W.Ba.)

During his first year as president, Mubarak struck a moderate note. There was no backing away from the peace with Israel and no loosening of the connection to the United States. By pursuing that steady course, he was able to prevent any delay in the return of the occupied Sinai to Egyptian sovereignty in April 1982. At the same time, Mubarak tried to contain the disaffections that had surfaced in the last year of Sādāt's era. He announced the end of the reign of the privileged minority that had dominated the invigorated private sector during the Sādāt years. He also moved immediately to soften the harsh edges of the authoritarianism of Free Officer rule. Unfortunately, Egypt's worsening economic problems did not lend themselves to rapid improvement. But in his very first speeches Mubarak did frankly and perceptively outline the shortcomings of the Egyptian economy.

These solid beginnings were undercut by the Israeli invasion of Lebanon in June 1982. In Egypt the invasion was seen as an Israeli attempt to destroy Palestinian nationalism, and Mubarak was accused by his detractors of allowing Israel to take advantage of Egypt's position of disengagement. Official relations with Israel were severely strained until Israel began its partial withdrawal from Lebanon in 1985. As a result of Mubarak's cautious policies, on the other hand, Egypt gradually was able to repair its relationships with most of the moderate Arab nations.

Within the country, opposition to a variety of political, economic, and social policies continued, chiefly among discontented labour and religious groups. The government contained labour strikes and other incidents of unrest and adopted several measures aimed at curbing a determined drive by Islāmic fundamentalists to destabilize the regime.

Egypt's economy suffered markedly from a sharp decline in oil prices in 1986 and was further weakened by a drop in the number of remittances from workers abroad. In spite of a rising burden of debt, the government continued to rely heavily on foreign economic aid.

For later developments in the history of Egypt, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 911, 912, 924, 962, 96/11, and 978. (R.W.Ba./D.H./Ed.)

BIBLIOGRAPHY

General works: RICHARD F. NYROP (ed.), *Egypt: A Country Study,* 4th ed. (1983), covering the history, society, economy, and politics of Egypt; SHIRLEY KAY, *The Egyptians: How They Live and Work* (1975), an introductory survey of Egypt's geography, history, government, and culture, as well as transportation; JASPER MORE, *The Land of Egypt* (1980), an illustrated general description of the country; and AHMED FAKHRY, *The Oases of Egypt,* 2 vol. (1973), a description of the oases of the Western Desert.

The land: W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography,* 7th rev. ed. (1978), basic geographical information; M.S. ABU AL-IZZ, *Landforms of Egypt,* trans. from Arabic (1971), a detailed outline of physiographic regionalization; MARTIN A.J. WILLIAMS and HUGUES FAURE (eds.), *The Sahara and the Nile: Quaternary Environments and Prehistoric Occupation in Northern Africa* (1980), a detailed geologic and anthropological study. Other specialized works include RUSHDI SAID, *The Geology of Egypt* (1962), and *The Geological Evolution of the River Nile* (1981); as well as TOM LITTLE, *High Dam at Aswan: The Subjugation of the Nile* (1965); and JULIAN RZÓSKA, *The Nile: Biology of an Ancient River* (1976), containing discussion of the biological effects of the Aswān High Dam. On plants and animals, see VIVI TÄCKHOLM, GUNNAR TÄCKHOLM, and MOHAMMED DRAR, *Flora of Egypt,* 4 vol. (1941–69, reprinted 1973), the standard work on the subject; RICHARD MEINERTZHAGEN, *Nicoll's Birds of Egypt,* 2 vol. (1930),

a primary source, copiously illustrated; and JOHN ANDERSON, WILLIAM E. DE WINTON, and GEORGE A. BOULENGER, *Zoology of Egypt*, 3 vol. in 4 (1898–1907, reprinted 1965), an authoritative and amply illustrated standard work. HENRY HABIB AYROUT, *The Fellaheen* (1945, reprinted 1981; originally published in French, 1938), contains observations on the customs, dress, and psychology of the Egyptian peasant; and HAMID AMMAR, *Growing Up in an Egyptian Village* (1954, reissued 1973), is an excellent and full account of village life in Egypt.

The people: ABBAS M. AMMAR, *The People of Sharqiya*, 2 vol. (1944), a physical anthropologist's description of the inhabitants of the eastern Delta; ROBERT A. FERNEA, *Nubians in Egypt: Peaceful People* (1973), an illustrated ethnographic essay; and ANWAR G. CHEJNE, *The Arabic Language: Its Role in History* (1969), a discussion of the background of classical Arabic and the dichotomy between it and the various dialects. WILLIAM H. WORRELL, *A Short Account of the Copts* (1945), is a concise study of the indigenous Christian population of Egypt. Other studies of religions of Egypt include OTTO F.A. MEINARDUS, *Christian Egypt, Ancient and Modern*, 2nd rev. ed. (1977), on the Christian communities; MORROE BERGER, *Islam in Egypt Today: Social and Political Aspects of Popular Religion* (1970); and G.H. JANSEN, *Militant Islam* (1979). For a popular introduction to the religions of Egypt, see VERONICA IONS, *Egyptian Mythology*, new ed. (1983).

Administrative and social conditions: HAROLD F. ALDERFER, M. FATHALLA EL KHATIB, and MOUSTAFA AHMED FAHMY, *Local Government in the United Arab Republic* (1963); MORROE BERGER, *Bureaucracy and Society in Modern Egypt* (1957, reprinted 1969); and P.J. VATIKIOTIS, *The Egyptian Army in Politics: Pattern for New Nations?* (1961, reprinted 1975). See also FRANK TACHAU (ed.), *Political Elites and Political Development in the Middle East* (1975); ENID HILL, *Mahkama!: Studies in the Egyptian Legal System: Courts & Crimes, Law & Society* (1979); JAMES B. MAYFIELD, *Local Institutions and Egyptian Rural Development* (1974); and HELMI R. TADROS, *Rural Resettlement in Egypt's Reclaimed Lands* (1978). On education, see AMIR BOKTOR, *The Development and Expansion of Education in the United Arab Republic* (1963), an important general survey; BAYARD DODGE, *Al-Azhar: A Millennium of Muslim Learning* (1961, reissued 1974); and GEORGIE D.M. HYDE, *Education in Modern Egypt: Ideals and Realities* (1978). Other works on social conditions include TOM LITTLE, *Modern Egypt* (1967), a study of social and political structures; PETER MANSFIELD, *Nasser's Egypt*, 2nd ed. (1969), a clear and orderly description of political, economic, and social changes in Egypt after 1952; UNNI WIKAN, *Life Among the Poor in Cairo* (1980; originally published in Norwegian, 1976); ABDEL R. OMRAN (ed.), *Egypt: Population Problems & Prospects* (1973); SAAD M. GADALLA, *Land Reform in Relation to Social Development, Egypt* (1962), and *Is There Hope?: Fertility and Family Planning in a Rural Egyptian Community* (1978); and ANDREA B. RUGH, *Family in Contemporary Egypt* (1984).

Economy: ROBERT L. TIGNOR, *State, Private Enterprise, and Economic Change in Egypt, 1918–1952* (1984); ROBERT MABRO, *The Egyptian Economy, 1952–1972* (1974); and ROBERT MABRO and SAMIR RADWAN, *The Industrialization of Egypt, 1939–1973: Policy and Performance* (1976); KASIM ALRIMAWI (QASIM RIMAWI), *The Challenge of Industrialization, Egypt* (1974); CHARLES ISSAWI, *Egypt in Revolution: An Economic Analysis* (1963, reprinted 1986); MOSTAFA H. NAGI, *Labor Force and Employment in Egypt: A Demographic and Socioeconomic Analysis* (1971); RASHED AL-BARAWY, *Economic Development in the United Arab Republic: Egypt* (1972); K.M. BARBOUR, *The Growth, Location, and Structure of Industry in Egypt* (1972); MAURICE GIRGIS, *Industrialization and Trade Patterns in Egypt* (1977); YUSUF J. AHMAD, *Absorptive Capacity of the Egyptian Economy: An Examination of Problems and Prospects* (1976); DAVID WILLIAM CARR, *Foreign Investment and Development in Egypt* (1979); KHALID IKRAM, *Egypt, Economic Management in a Period of Transition* (1980); and JOHN WATERBURY, *The Egypt of Nasser and Sadat: The Political Economy of Two Regimes* (1983).

Cultural life: MUSTAFA HABIB (ed.), *Cultural Life in the United Arab Republic* (1968); ALBERT HOURANI, *Arabic Thought in the Liberal Age, 1798–1939* (1962, reissued 1983), a study of the interaction of Western and indigenous culture in its historical context; JACOB M. LANDAU, *Studies in the Arab Theatre and Cinema* (1958); ABD AL-MONEM ISMAIL, *Drama and Society in Contemporary Egypt* (1967); FAROUK ABDEL WAHAB (comp.), *Modern Egyptian Drama* (1974); ABDEL-AZIZ ABDEL-MEGUID, *The Modern Arabic Short Story: Its Emergence, Development, and Form* (1950?); HAMDI SAKKUT, *The Egyptian Novel and Its Main Trends from 1913 to 1952* (1971); and HILARY KILPATRICK, *The Modern Egyptian Novel: A Study in Social Criticism* (1974). Other studies include MOUHAN A. KHOURI, *Poetry and the Making of Modern Egypt, 1882–1922* (1971); YVES THORAVAL, *Regards sur le cinéma Égyptien* (1975), on the Egyptian cinema; and PIERRE DU BOURGUET, *Coptic Art* (1971, originally published in French, 1968).

History: (Ancient Egypt): The most detailed presentation of Egyptian history, with full bibliographies arranged by subject, is the multivolume *Cambridge Ancient History,* though volumes 1 and 2 no longer reflect current knowledge. WOLFGANG HELCK, EBERHARD OTTO, and WOLHART WESTENDORF (eds.), *Lexikon der Ägyptologie* (1975–), is the basic reference work in Egyptology, of which 5 volumes had appeared by 1986, with further parts published in fascicles. MICHAEL A. HOFFMAN, *Egypt Before the Pharaohs: The Prehistoric Foundations of Egyptian Civilization* (1979, reissued 1984), is a comprehensive general work on prehistory; while LECH KRZYŻANIAK, *Early Farming Cultures on the Lower Nile: The Predynastic Period in Egypt* (1977), focuses on the transition to agriculture and on Lower Egypt. General studies include CYRIL ALDRED, *The Egyptians,* rev. ed. (1984); and JOHN RUFFLE, *Heritage of the Pharaohs: An Introduction to Egyptian Archaeology* (1977); as well as other works cited below under the specific periods on which they focus. General histories include B.G. TRIGGER *et al., Ancient Egypt: A Social History* (1983), containing four essays on the main periods, concentrating on relations with Africa, and including valuable bibliographies; and SIR ALAN GARDINER, *Egypt of the Pharaohs* (1961), a personal history, notable for the use made of ancient Egyptian texts. WILLIAM W. HALLO and WILLIAM KELLY SIMPSON, *The Ancient Near East: A History* (1971), is a reliable brief introduction; and ÉTIENNE DRIOTON and JACQUES VANDIER, *L'Égypte: des origines à la conquête d'Alexandre,* 4th ed. (1962, reprinted 1984), remains valuable for its critical discussions. JOHN A. WILSON, *The Burden of Egypt: An Interpretation of Ancient Egyptian Culture* (1951, reprinted 1965), is a selective historical study. WILLIAM C. HAYES, *The Scepter of Egypt,* 2 vol. (1953–59), is a detailed cultural history of Egypt to the end of the 20th dynasty, based on the collections in the Metropolitan Museum of Art, New York City. WOLFGANG HELCK, *Geschichte des alten Ägypten* (1968, reprinted 1981), is still the best general history; his *Beziehungen Ägyptens zu Vorderasien im 3. und 2. Jahrtausend v. Chr.,* 2nd ed. (1971), is the fundamental work on foreign relations, and his *Wirtschaftsgeschichte des Alten Ägypten im 3. und 2. Jahrtausend vor Chr.* (1975) covers institutions and economics. ROLF KRAUSS, *Sothis- und Monddaten: Studien zur astronomischen und technischen Chronologie Altägyptens* (1985), is a vital chronological study for the 2nd and 1st millennia BC; its dates are adopted in this article with minor variations. JOHN BAINES and JAROMÍR MÁLEK, *Atlas of Ancient Egypt* (1980), is a concise, geographically oriented survey. HERMANN KEES, *Ancient Egypt: A Cultural Topography* (1961, reprinted 1977; originally published in German, 2nd ed., 1958; 3rd German ed., 1977), studies a number of major sites in depth. KARL W. BUTZER, *Early Hydraulic Civilization in Egypt: A Study in Cultural Ecology* (1976), is a useful discussion of geographic and environmental conditions and their relation to the development of ancient Egyptian civilization. CLAUDE VANDERSLEYEN *et al., Das alte Ägypten* (1975), is the most comprehensive survey of Egyptian art. W. STEVENSON SMITH, *The Art and Architecture of Ancient Egypt,* rev. ed., edited by WILLIAM KELLY SIMPSON (1981), is an excellent general account; and for the Old Kingdom, Smith's *History of Egyptian Sculpture and Painting in the Old Kingdom,* 2nd ed. (1949), is still a fundamental source. MIRIAM LICHTHEIM, *Ancient Egyptian Literature: A Book of Readings,* 3 vol. (1973–80), offers an excellent collection of texts in translation, covering the Old, Middle, and New Kingdoms and the Late Period. A smaller selection of readings is available in WILLIAM KELLY SIMPSON (ed.), *The Literature of Ancient Egypt: An Anthology of Stories, Instructions, and Poetry,* new ed. (1973); while JAMES B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament,* 3rd ed. (1969), contains a wide selection of Egyptian material in translation. Studies of administration include KLAUS BAER, *Rank and Title in the Old Kingdom* (1960, reprinted 1974); to which NIGEL STRUDWICK, *The Administration of Egypt in the Old Kingdom: The Highest Titles and Their Holders* (1985), adds a vast amount of detail. WOLFGANG HELCK, *Zur Verwaltung des Mittleren und Neuen Reichs* (1958), with a separately published index volume (1975), is the basic work on the succeeding periods.

(Egypt from the 18th dynasty to 332 BC): The rise of the New Kingdom is treated in JÜRGEN VON BECKERATH, *Untersuchungen zur politischen Geschichte der Zweiten Zwischenzeit in Ägypten* (1964). DONALD B. REDFORD, *History and Chronology of the Eighteenth Dynasty of Egypt: Seven Studies* (1967), includes a reevaluation of Hatshepsut. An informative account of the New Kingdom empire at its height is ELIZABETH RIEFSTAHL, *Thebes in the Time of Amunhotep III* (1964, reprinted 1971). For the controversial Amarna period, ROLF KRAUSS, *Das Ende der Amarnazeit: Beitr. zur Geschichte u. Chronologie d. Neuen Reiches* (1978); and DONALD B. REDFORD, *Akhenaten, the Heretic King* (1984), offer strongly contrasting interpre-

tations. CYRIL ALDRED, *Akhenaten and Nefertiti* (1973), is a good collection of the artistic evidence for the period. For the Ramesside period, K.A. KITCHEN, *Pharaoh Triumphant: The Life and Times of Ramesses II King of Egypt* (1982), sets its subject in context, presenting the New Kingdom in general as well as Ramses' own reign. EDWARD F. WENTE, *Late Ramesside Letters* (1967), deals with material from the end of the same period. For the economy of this time, see the major work of J.J. JANSSEN, *Commodity Prices from the Ramessid Period: An Economic Study of the Village of Necropolis Workmen at Thebes* (1975). JOHN ROMER, *Ancient Lives: Daily Life in Egypt of the Pharaohs* (1984), presents the life of the same community. T.G.H. JAMES, *Pharaoh's People: Scenes from Life in Imperial Egypt* (1984), is concerned with life-styles of higher ranks of society in the same general period. K.A. KITCHEN, *The Third Intermediate Period in Egypt (1100–650 B.C.)*, 2nd rev. ed. (1986), is the basic work on the period. HERMANN KEES, *Das Priestertum im ägyptischen Staat, vom neuen Reich bis zur Spätzeit* (1953), with an index volume, *Indices und Nachträge* (1958), is a comprehensive analysis of the Egyptian priesthoods. This fundamental institution of the Late Period is also valuably treated in SERGE SAUNERON, *The Priests of Ancient Egypt* (1960, reprinted 1980; originally published in French, 1957). On the period from the Saite 26th dynasty until Alexander the Great, see FRIEDRICH K. KIENITZ, *Die politische Geschichte Ägyptens vom 7. bis 4. Jahrhundert vor der Zeitwende* (1953), based on both Egyptian and classical sources. ALAN B. LLOYD, *Herodotus, Book II*, 2 vol. (1975–76), contains much material on the Late Period.

(*Hellenistic and Roman Egypt*): On the period in general, see HAROLD I. BELL, *Egypt, from Alexander the Great to the Arab Conquest: A Study in the Diffusion and Decay of Hellenism* (1948, reprinted 1980); and ALAN K. BOWMAN, *Egypt After the Pharaohs, 332 B.C.–A.D. 642: From Alexander to the Arab Conquest* (1986). The basic general works on the papyri are L. MITTEIS and U. WILCKEN, *Grundzüge und Chrestomathie der Papyruskunde*, 2 vol. in 4 (1912, reprinted 1963); and E.G. TURNER, *Greek Papyri: An Introduction* (1968, reissued 1980), with its illustrated companion, *Greek Manuscripts of the Ancient World* (1971). On Ptolemaic Egypt, see DOROTHY J. CRAWFORD, *Kerkeosiris: An Egyptian Village in the Ptolemaic Period* (1971); P.M. FRASER, *Ptolemaic Alexandria*, 3 vol. (1972); J. GRAFTON MILNE, *A History of Egypt Under Roman Rule*, 3rd rev. ed. (1924); ORSOLINA MONTEVECCHI, *La papirologia* (1973); ALAN E. SAMUEL, *From Athens to Alexandria: Hellenism and Social Goals in Ptolemaic Egypt* (1983); NAPHTALI LEWIS, *Greeks in Ptolemaic Egypt: Case Studies in the Social History of the Hellenistic World* (1986); E.E. RICE, *The Grand Procession of Ptolemy Philadelphus* (1983); M. ROSTOVTZEFF, *The Social & Economic History of the Hellenistic World*, 3 vol. (1941, reprinted with corrections 1972); and SARAH B. POMEROY, *Women in Hellenistic Egypt: From Alexander to Cleopatra* (1984). On Roman Egypt, see A.C. JOHNSON, *Roman Egypt to the Reign of Diocletian*, vol. 2 in TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome*, 6 vol. (1933–40, reprinted 1975); A.H.M. JONES, *The Cities of the Eastern Roman Provinces*, 2nd ed. (1971, reprinted 1983); and NAPHTALI LEWIS, *Life in Egypt Under Roman Rule* (1983). On Byzantine Egypt, see ALFRED J. BUTLER, *The Arab Conquest of Egypt and the Last Thirty Years of the Roman Dominion*, 2nd ed., revised by P.M. FRASER (1978); EDWARD ROCHIE HARDY, *The Large Estates of Byzantine Egypt* (1931, reprinted 1968), and *Christian Egypt: Church and People: Christianity and Nationalism in the Patriarchate of Alexandria* (1952); ALLAN CHESTER JOHNSON and LOUIS C. WEST, *Byzantine Egypt: Economic Studies* (1949, reprinted 1967); and COLIN H. ROBERTS, *Manuscript, Society, and Belief in Early Christian Egypt* (1979).

(*Egypt from c. 630 to c. 1800*): Two standard works that survey medieval Egyptian history as a whole are STANLEY LANE-POOLE, *A History of Egypt in the Middle Ages*, 4th ed. (1968); and GASTON WIET, *L'Égypte arabe de la conquête arabe à la conquête ottomane, 642–1517 de l'ère chrétienne*, vol. 4 in GABRIEL HANOTAUX, *Histoire de la nation égyptienne*, 7 vol. (1931–40). Each of these is outdated in many respects, but each presents an accurate summary of the political history of the period, based on primary Arabic sources; also, both are strong on Egyptian architecture as an insight into political, social, and economic history. A valuable later reference source with comprehensive coverage of the period is JOAN WUCHER KING, *Historical Dictionary of Egypt* (1984). For the economic history, see SUBHI LABIB, *Handelsgeschichte Ägyptens im Spätmittelalter, 1171–1517* (1965); Labib has summarized this book in English in the form of an article, "Egyptian Commercial Policy in the Middle Ages," in *Studies in the Economic History of the Middle East: From the Rise of Islam to the Present Day*, edited by M.A. COOK, pp. 63–77 (1970). ELIYAHU ASHTOR, *A Social and Economic History of the Near East in the Middle Ages* (1976), and *Levant Trade in the Later Middle Ages* (1983), are also important. AZIZ S. ATIYA, *A History of Eastern Christianity* (1968, reissued 1980), is authoritative for Coptic history. For the beginnings of Muslim Egypt, see FRANCESCO GABRIELI, *Muhammad and the Conquests of Islam* (1968, originally published in Italian, 1967), for the conquest of Egypt; and DANIEL C. DENNETT, *Conversion and the Poll Tax in Early Islam* (1950), for Muslim tax policy in Egypt. For the Ṭūlūnids, see ZAKY MOHAMED HASAN, *Les Tulunides* (1933). Fāṭimid studies have been transformed by S.D. GOITEIN, *A Mediterranean Society: The Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza* (1968–), of which four volumes had appeared by 1987. Three articles by HAMILTON A.R. GIBB are definitive for Egypt under the Ayyūbids and during the Crusades, all published in *A History of the Crusades*, ed. by KENNETH M. SETTON, 2nd ed., 5 vol. (1958–85): "The Caliphate and the Arab States," 1:81–98; "The Rise of Saladin, 1169–1189," 1:563–589; and "The Aiyūbids," 2:693–714. See also R. STEPHEN HUMPHREYS, *From Saladin to the Mongols: The Ayyubids of Damascus, 1193–1260* (1977). For Mamlūk and Ottoman Egypt, see F.R.C. BAGLEY (ed. and trans.), *The Last Great Muslim Empires*, vol. 3 in *The Muslim World: A Historical Survey*, 3 vol. (1960–69, originally published in German, 1952–59). An account of the early Mamlūk state is found in ROBERT IRWIN, *The Middle East in the Middle Ages: The Early Mamluk Sultanate, 1250–1382* (1986); and for the Ottoman period alone, see STANFORD J. SHAW, *The Financial and Administrative Organization and Development of Ottoman Egypt, 1517–1798* (1962).

(*Egypt since 1800*): EDWARD WILLIAM LANE, *An Account of the Manners and Customs of the Modern Egyptians*, 2 vol. (1836, reissued in 1 vol., 1973), is a classic study of everyday life during the second quarter of the 19th century. An analysis of the political developments of the period is offered in F. ROBERT HUNTER, *Egypt Under the Khedives, 1805–1879: From Household Government to Modern Bureaucracy* (1984). JAMAL M. AHMED, *The Intellectual Origins of Egyptian Nationalism* (1960, reissued 1968), is particularly concerned with the nationalists of the period from 1892 to 1914. Other useful works are GABRIEL BAER, *A History of Landownership in Modern Egypt, 1800–1950* (1962); P.M. HOLT, *Egypt and the Fertile Crescent, 1516–1922* (1966), and P.M. HOLT (ed.), *Political and Social Change in Modern Egypt* (1968); JACOB M. LANDAU, *Parliaments and Parties in Egypt* (1953, reissued 1979); HELEN ANNE B. RIVLIN, *The Agricultural Policy of Muḥammad ʿAlī in Egypt* (1961); AFAF LUFTI AL-SAYYID MARSOT, *Egypt in the Reign of Muhammad Ali* (1984), a sturdy defense by an Egyptian author; ROBERT L. TIGNOR, *Modernization and British Colonial Rule in Egypt, 1882–1914* (1966); and NADAV SAFRAN, *Egypt in Search of Political Community: An Analysis of the Intellectual and Political Evolution of Egypt, 1804–1952* (1961, reprinted 1981). P.J. VATIKIOTIS, *Nasser and His Generation* (1978), offers a fine biography, especially for the years between 1930 and 1952; RAYMOND W. BAKER, *Egypt's Uncertain Revolution Under Nasser and Sadat* (1978), analyzes the effect of the Egyptian revolution on Egyptian society; DAVID HIRST and IRENE BEESON, *Sadat* (1981), is an early assessment of the Sādāt years; RAYMOND A. HINNEBUSCH, JR., *Egyptian Politics Under Sadat: The Post-Populist Development of an Authoritarian-Modernizing State* (1985), is an interesting study; DEREK HOPWOOD, *Egypt. Politics and Society, 1945–1984*, 2nd ed. (1985), is a general comprehensive introduction; and P.J. VATIKIOTIS, *The History of Egypt*, 3rd ed. (1985), together with AFAF LUFTI AL-SAYYID MARSOT, *A Short History of Modern Egypt* (1985), are especially valuable for their analyses of the post-Sādāt period.

(L.S.El.H./Ma.J./C.G.S./D.H./
J.R.Ba./A.K.B./D.S.Ri.)

Ancient Egyptian Arts and Architecture

In the general tradition of the visual arts of the West, ancient Egypt represents a source of form and technique dating back to the early 3rd millennium BC. For the purposes of definition ancient Egyptian is essentially coterminous with dynastic Egyptian, the dynastic structure of Egyptian history, artificial though it may partly be, providing a convenient chronological framework. The distinctive periods are: Early Dynastic (1st–3rd dynasties, *c.* 2925–*c.* 2575 BC); Old Kingdom (4th–8th dynasties, *c.* 2575–*c.* 2130 BC); First Intermediate (9th–11th dynasties, *c.* 2130–1939 BC); Middle Kingdom (12th–14th dynasties, 1938–*c.* 1600? BC); Second Intermediate (15th–17th dynasties, *c.* 1630–1540 BC); New Kingdom (18th–20th dynasties, 1539–1075 BC); Third Intermediate (21st–25th dynasties, *c.* 1075–656 BC); and Late Dynastic (26th–31st dynasties, 664–332 BC).

Geographical factors were predominant in forming the particular character of Egyptian art. By providing Egypt with the most predictable agricultural system in the ancient world, the Nile afforded a stability of life in which arts and crafts readily flourished. Equally, the deserts and the sea, which protected Egypt on all sides, contributed to this stability by discouraging serious invasion for almost 2,000 years. The desert hills were also rich in minerals and fine stones, ready to be exploited by artists and craftsmen. Only good wood was lacking, and the need for it led the Egyptians to undertake foreign expeditions to Lebanon, to Somalia, and, through intermediaries, to tropical Africa. In general, the search for useful and precious materials determined the direction of foreign policy and the establishment of trade routes and led ultimately to the enrichment of Egyptian material culture. For further treatment, see EGYPT; MIDDLE EASTERN RELIGIONS, ANCIENT.

This article is divided into the following sections:

Predynastic Period

The term predynastic denotes the period of emerging cultures that preceded the establishment of the 1st dynasty in Egypt. In the late 5th millennium BC there began to emerge patterns of civilization that displayed characteristics deserving to be called Egyptian. The accepted sequence of predynastic cultures is based on the excavations of Sir Flinders Petrie at Naqādah, at al-ʿĀmirah (el-ʿÂmra), and at al-Jazīrah (el-Gezira). Another somewhat earlier stage of predynastic culture has been identified at al-Badārī in Upper Egypt.

From graves at al-Badārī, Dayr Tasa, and al-Mustaqiddah evidence of a relatively rich and developed artistic and industrial culture has been retrieved. Pottery of a fine red polished ware with blackened tops already shows distinctive Egyptian shapes. Copper was worked into small ornaments, and beads of steatite (soapstone) show traces of primitive glazing. Subsequently in the Naqādah I and Naqādah II stages predynastic civilization developed steadily. Pottery remains the distinctive product, showing refinement of technique and the development of adventurous decoration. Shapes already found in Badarian graves were produced in Naqādah I with superior skill and decorated with geometric designs of white-filled lines and even simple representations of animals. Later new clays were exploited, and fine buff-coloured wares were decorated in purple pigment with scenes of ships, figures, and a wide variety of symbols.

Refinement of pottery

The working of hard stones also began in earnest in the later Predynastic Period. At first craftsmen were devoted to the fashioning of fine vessels and to the making of jewelry incorporating semiprecious stones.

Sculpture found its best beginnings not so much in representations of the human form (although figurines, mostly female, were made from Badarian times) as in the carving of small animal figures and the making of schist (slate) palettes (intended originally for the preparation of eye paint). The Hunters and Battlefield palettes (British Museum; part of the former in the Louvre; part of the latter in the Ashmolean, Oxford) show two-dimensional representation—a convention that was to last 3,000 years.

The schist palettes

The basic techniques of two-dimensional art—drawing and painting—are exemplified in Upper Egyptian rock drawings and in the painted tomb at Hierakonpolis, now destroyed. Scenes of animals, boats, and hunting, the common subjects of rock drawings, were more finely executed in paint in the tomb, and additional themes, probably of conquest, presaged those found in dynastic art.

Dynastic Egypt

Evidence suggests that the unification of Upper and Lower Egypt drew together the various threads of what was to become the rich tapestry of Egyptian culture and started the intricate weave on the loom of time. Many of the new artistic developments undoubtedly can be traced back to the Naqādah II period; but the abundant evidence from the great tombs of the 1st dynasty at Abydos and Ṣaqqārah far outweighs what was found in the modest burials of earlier times. The impression is certainly one of an extraordinary efflorescence of civilization. Conquest, implicit in unification, is dramatically characterized in the scenes shown on the Narmer Palette (Egyptian Museum, Cairo), where Narmer, probably the founding king of dynastic Egypt, and better known as Menes, is depicted as the triumphant ruler (Figure 1).

The Narmer representations display much of what is typical of Egyptian art of the Dynastic Period. Here is the characteristic image of the king smiting his enemy, depicted with the conventions that distinguish Egyptian two-dimensional art. The head is shown in profile, but the eye in full; the shoulders are frontally represented, while the torso is at three-quarters view; the legs again are in profile. To show as much detail as possible was the principal intention of the artist—to show what he knew was

Conventions of two-dimensional representation

there, not simply what he could see from one viewpoint. Further conventions, well established by the 4th dynasty, included the showing of both hands and feet, right and left, without distinction. Scenes were set on baselines, and the events were placed in sequence, usually from right to left. Unity in a scene was provided by the focal figure of the most important person, the king or tomb owner. Relative size established importance: the ruler dwarfed the high official, while the tomb owner dwarfed his wife and, still more so, his children.

The canon of proportion

Conservatism in artistic matters was nurtured by a relative coherence of culture, strengthened by a vigorous tradition of scribal training, and tempered by a canon of proportion for the representation of the human figure. In the Old Kingdom, walls prepared for decoration were marked out with red horizontal guidelines; in later times vertical lines were added. During much of the Dynastic Period a grid of 18 rows of squares was used to contain the standing figure of a man; from the 26th dynasty, 21 rows of squares were used for the same purpose. At different periods, variations in the placing of specific bodily features produced interesting and subtle nuances. During the so-called Amarna period a distinctive reappraisal of the canon took place. The full range of changes and the many variants still remain to be studied, but it is clear that the basic canon lay deeply rooted in the training of the Egyptian artist.

Figure 1: Slate Narmer Palette, from Hierakonpolis, beginning of 1st dynasty, *c.* 2925 BC. In the Egyptian Museum, Cairo. Height 63.5 cm. (Left) Obverse, divided into three pictorial strips: the king, wearing the crown of Lower Egypt, shown on his way to witness the execution of fettered enemies; two bearded men leading two fabulous animals, perhaps symbolizing the unification of Upper and Lower Egypt; and the king in the form of a wild ox attacking a fortified settlement. (Right) Reverse, showing a victory motive: King Narmer, wearing the crown of Upper Egypt, striking down an enemy held by the hair.

Hirmer Fotoarchiv, Munchen

ARCHITECTURE

The two principal building materials used in ancient Egypt were unbaked mud brick and stone. From the Old Kingdom onward stone was generally used for tombs—the eternal dwellings of the dead—and for temples—the eternal houses of the gods. Mud brick remained the domestic material, used even for royal palaces; it was also used for fortresses, the great walls of temple precincts and towns, and for subsidiary buildings in temple complexes.

Most ancient Egyptian towns have been lost because they were situated in the cultivated and flooded area of the Nile Valley; many temples and tombs have survived because they were built on ground unaffected by the Nile flood. Any survey of Egyptian architecture will in consequence be weighted in favour of funerary and religious buildings. Yet the dry, hot climate of Egypt has allowed some mud brick structures to survive where they have escaped the destructive effects of water or man.

Tomb architecture. Mortuary architecture in Egypt was highly developed and often grandiose. The tomb was not simply a place in which a corpse might be protected from desecration. It was the home of the deceased, provided with material objects to ensure continued existence after death. Part of the tomb might reproduce symbolically the earthly dwelling of the dead person; it might be decorated with scenes that would enable the individual to pursue magically an afterlife suitable and similar to his worldly existence. For a king the expectations were quite different; for him the tomb became the vehicle whereby he might achieve his exclusive destiny with the gods in a celestial afterlife.

The tomb as a home

Most tombs comprised two principal parts, the burial chamber (the tomb proper) and the chapel, in which offerings for the deceased could be made. In royal burials the chapel rapidly developed into a temple, which in later times was usually built separately and at some distance from the tomb. In the following discussion, funerary temples built separately will be discussed with temples in general and not as part of the funerary complex.

Royal tombs. In the earliest dynasties the tombs of kings and high officials were made of mud brick and of such similar size that it is difficult to distinguish between them. It is now generally thought that the tombs at Abydos were royal, whereas those at Ṣaqqārah were noble. The latter, better preserved than the former, reveal rectangular superstructures, called mastabas (see below), with sides constructed in the form of paneled niches painted white and decorated with elaborate "matting" designs.

These great superstructures contained many storage chambers stocked with food and equipment for the deceased, who lay in a rectangular burial chamber below ground. Also within the superstructure, but not always clearly evident, was a low mound of earth, possibly representing the primitive grave of earlier times. Sometimes this concealed mound was a low, stepped structure, perhaps the precursor of the first great building constructed of stone in Egypt.

The Step Pyramid of Djoser

The Step Pyramid of Djoser, second king of the 3rd dynasty, was built within a vast enclosure on a commanding site at Ṣaqqārah overlooking the city of Memphis. A high royal official, Imhotep, has traditionally been credited with the design and with the decision to use quarried stone. This first essay in stone is remarkable for its design of six superposed stages of diminishing size, and also for its huge enclosure (1,784 × 909 feet [544 × 277 metres]) surrounded by a paneled wall faced with fine limestone and containing a series of "mock" buildings that probably represent structures associated with the palace in Memphis. There the Egyptian stonemasons made their earliest architectural innovations, using stone to reproduce the forms of primitive wood and brick buildings. Fine reliefs of the king and elaborate wall "hangings" in glazed tiles in parts of the subterranean complexes are among the innovations found in this remarkable monument.

The Great Pyramid of Khufu

For the Old Kingdom the most characteristic form of tomb building was the true pyramid, the finest example of which is the Great Pyramid of King Khufu (Cheops) of the 4th dynasty at al-Jīzah (Giza; Figure 2). The form itself reached its maturity in the reign of Snefru, father of Khufu. Subsequently only the pyramid of Khafre (Chephren), Khufu's successor, approached the size and perfection of the Great Pyramid. The simple measurements of the Great Pyramid indicate very adequately its scale, monumentality, and precision: its sides are 755.43 feet (230.26 metres; north), 756.08 feet (230.45 metres; south), 755.88 feet (230.39 metres; east), 755.77 feet (230.36 metres; west); its orientation on the cardinal points is almost exact; its height upon completion was 481.4 feet (146.7 metres); its area at base is just over 13 acres (5.3 hectares). Other features in its construction contribute substantially to its remarkable character: the lofty, corbeled Grand Gallery and the granite-built King's Chamber with five relieving compartments (empty rooms for reducing pressure) above.

The pyramid formed the focal point of a group of buildings that constituted the funerary complex of a king. Two temples linked by a causeway were essential components. The valley temple, built on the edge of the desert escarp-

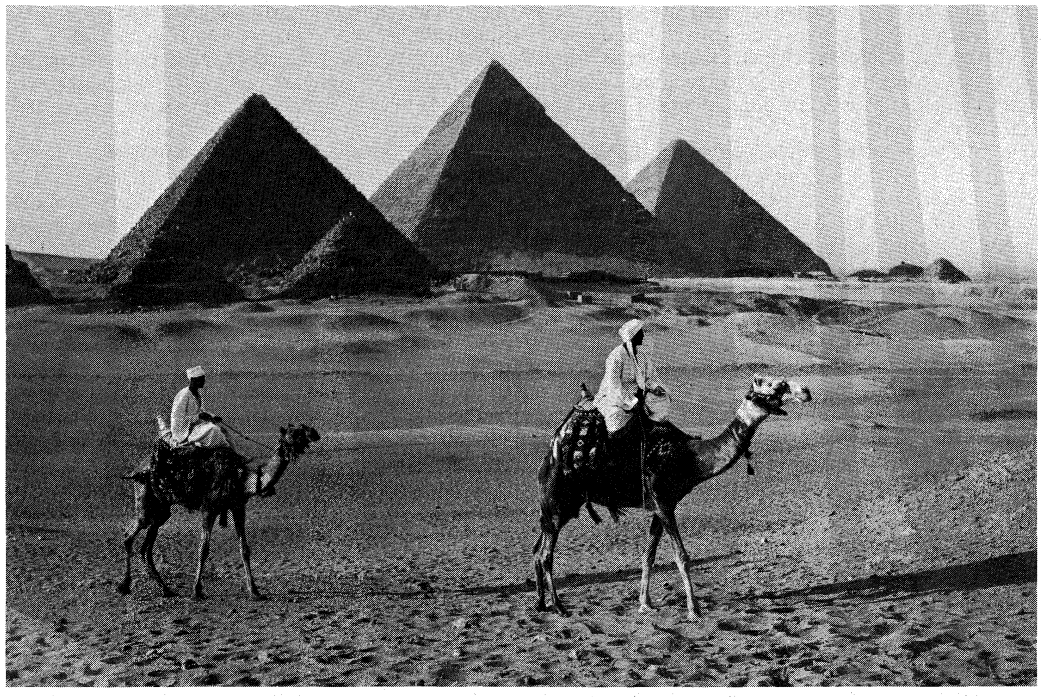


Figure 2: The Pyramids of Giza with the Great Pyramid of King Khufu (Cheops), 4th dynasty (*c.* 2575–*c.* 2465 BC), to the right.
Ray Manley—Shostal Assoc.

ment, was the place of reception for the royal body. The most striking valley temple is that of Khafre, a structure of massive granite blocks with huge alabaster flooring slabs, starkly simple but immensely effective. The best preserved causeway serves the pyramid of King Unas of the 5th dynasty; it contains low-relief wall decorations and a ceiling adorned with stars. The pyramid temple of Unas is distinguished by the extensive use of granite for architectural elements, including doorways and splendid monolithic columns with palm capitals.

The pyramids built for the later kings of the Old Kingdom and most kings of the Middle Kingdom were comparatively poor in size, construction, and materials. The tomb of King Mentuhotep II of the 11th dynasty is, however, of exceptional interest. The tomb complex at Dayr al-Baḥrī was once thought to have contained a pyramid, but excavations between 1966 and 1971 have shown that the hypothetical reconstructions were misconceived. Its essential components were a rectangular structure, a series of pillared ambulatories, an open court, and a hypostyle hall tucked into the cliffs.

The monumentality of the pyramid made it not only a potent symbol of royal power but also an obvious target for tomb robbers. During the New Kingdom the wish to halt the robbing and desecration of royal tombs led to their being sited together in a remote valley at Thebes, dominated by a peak that itself resembled a pyramid.

The Valley of the Kings

There, in the Valley of the Kings, tombs were carved deep into the limestone with no outward structure and marked only by a doorway carved in the rock face. They had no common plan, but most consisted of a series of corridors opening out at intervals to form rooms and ending in a large burial chamber deep in the mountain. The finest of the tombs is that of Seti I, second king of the 19th dynasty; it extends 328 feet (100 metres) into the mountain and contains a spectacular burial chamber, the barrel-shaped roof of which represents the vault of heaven.

After the abandonment of the valley at the end of the 20th dynasty, kings of the subsequent two dynasties were buried in very simple tombs within the temple enclosure of the delta city of Tanis. No later royal tombs have ever been identified.

Private tombs. A major distinction between royal and nonroyal tombs lies in the provision of arrangements for the funerary cult of the deceased. The evidence available from the 3rd dynasty onward makes it clear that king and commoner had quite different expectations. In nonroyal tombs a chapel was provided that included a formal tablet or stela on which the deceased was shown seated at a table of offerings. The earliest examples are simple and architecturally undemanding; later a suitable room, the tomb-chapel, was provided for the stela (now incorporated in a false door) in the tomb superstructure, or mastaba.

The mastaba

The term mastaba (Arabic: "bench") was first used archaeologically in the 19th century by workmen on Auguste Mariette's excavation at Ṣaqqārah to describe the rectangular, flat-topped stone superstructures of tombs. Subsequently, mastaba was also used for mud brick superstructures.

In the great cemeteries of the Old Kingdom, changes in size, internal arrangements, and groupings of the burials of nobles indicate the vicissitudes of nonroyal posthumous expectations. In the 3rd dynasty at Ṣaqqārah the most important private burials were at some distance from the step pyramids of Djoser and Sekhemkhet. Their large superstructures incorporated offering niches that were to develop into chapels (as in the tomb of Khabausokar) and corridors that could accommodate paintings of equipment for the afterlife and niches to hold carved representations of the deceased owner (as in the tomb of Hesire). During the 4th dynasty the stone mastabas of the Giza pyramid field were regularly laid out near the pyramids, and, although smaller than those at Ṣaqqārah, they show the true start of the exploitation of space within the superstructure. The niche chapel became a room for the false door and offering table, and there might also be rooms containing scenes of offering and of daily activities.

Nothing indicates more clearly the relaxation of royal authority in the later Old Kingdom than the size and decoration of the mastabas at Ṣaqqārah and Abusīr. Externally they were still rectangular structures, occasionally with a low wall establishing a precinct (as in the tomb of Mereruka). The full exploitation of internal space in the great mastabas at Abusīr (that of Ptahshepses) and Ṣaqqārah (that of Ti and the double mastaba of Akhtihotep and Ptahhotep) made ample room available for the receipt of offerings and for the representation of the milieu in which the dead owner might expect to spend his afterlife. In the mastaba of Mereruka, a vizier of Teti, first king of the 6th dynasty, there were 21 rooms for his own funerary purposes, with six for his wife and five for his son.

Exploitation of internal space

Contemporaneously, the provincial colleagues of the

Memphite nobles developed quite different tombs in Middle and Upper Egypt. Tomb chapels were excavated into the rock of the cliffs overlooking the Nile. Rock-cut tombs subsequently were to become the most common kind of private tomb, although mastabas were built in the royal cemeteries of the 12th dynasty.

Rock-cut tombs

Most rock-cut tombs were fairly simple single chambers serving all the functions of the multiplicity of rooms in a mastaba. Some, however, were excavated with considerable architectural pretensions. At Aswān huge halls, often connecting to form labyrinthine complexes, were partly formal, with columns carefully cut from the rock, and partly rough-hewn. Chapels with false doors were carved out within the halls. In some cases the facades were monumental, with porticoes and inscriptions.

At Beni Hasan the local nobles during the Middle Kingdom cut large and precise tomb chambers in the limestone cliffs. Architectural features—columns, barrel roofs, and porticoes, all carved from the rock—provided fine settings for painted mural decorations. The tombs of Khnumhotep and Amenemhet are outstanding examples of fine design impeccably executed.

The most famous rock-cut private tombs are those of the New Kingdom at Thebes, their fame resting, above all, on their mural decoration. As elsewhere the excavated chambers are the tomb-chapels, mostly taking a simple T-form, in which the crossbar of the T represents the entrance hall, and the upright stroke of the T is the chapel proper. Some of the more important tombs (Rekhmire, Ramose) have open courts before their unelaborate facades and some striking internal features, but most are small in comparison with those of earlier times.

A separate tradition of private tomb design was developed for important officials at Ṣaqqārah in the New Kingdom. Open courts, constructed offering chapels, and elaborate subterranean suites of rooms characterize these Memphite tombs. The tomb for Horemheb, a military commander who became the last king of the 18th dynasty, has remarkable relief decoration. The tomb of Tia (a sister of the 19th-dynasty king Ramses II) has a small pyramid behind the chapel.

Temple architecture. Two principal kinds of temple can be distinguished—cult temples and funerary or mortuary temples. The former accommodated the images of deities, the recipients of the daily cult; the latter were the shrines for the funerary cults of dead kings.

Cult temples. It is generally thought that the Egyptian temple of the Dynastic Period owed most to the cult of the sun god Re at Heliopolis. The temple of Re, however, was probably open in plan and lacked a shrine. Sun temples were unique among cult temples; worship was centred on a cult object, the *benben,* which was a squat obelisk placed in full sunlight. Among the few temples surviving from the Old Kingdom are sun temples of the 5th-dynasty kings at Abū Jirāb (Abu Gurab). That of Neuserre reveals the essential layout: a reception pavilion at the desert edge connected by a covered corridor on a causeway to the open court of the temple high on the desert, within which stood the *benben* of limestone and a huge alabaster altar. Fine reliefs embellished the covered corridor and also corridors on two sides of the court.

Sun temples

The cult temple achieved its most highly developed form in the great sanctuaries erected over many centuries at Thebes. Architecturally the most satisfying, and certainly the most beautiful, is the Luxor Temple, started by Amenhotep III of the 18th dynasty. The original design consists of an imposing open court with colonnades of graceful lotus columns, a smaller offering hall, a shrine for the ceremonial boat of the god, an inner sanctuary for the cult image, and a room in which the divine birth of the king was celebrated. The approach to the temple was made by a colonnade of huge columns with open papyrus-flower capitals, planned by Amenhotep III but decorated with fascinating processional reliefs under Tutankhamen and Horemheb. Later Ramses II built a wide court before the colonnade and two great pylons to form a new entrance.

Temple at Luxor

The necessary elements of an Egyptian temple, most of which can be seen at Luxor, are the following: an approach avenue of sphinxes leading to the great double-towered pylon entrance fitted with flagpoles and pennants; before the pylon a pair of obelisks and colossal statues of the king; within the pylon a court leading to a pillared hall, the hypostyle, beyond which might come a further, smaller hall where offerings could be prepared; and at the heart of the temple, the shrine for the cult image. In addition, there were storage chambers for temple equipment and sometimes a crypt. Outside the main temple building was a lake, or at least a well, for the water needed in the rituals; in later times there might also be a birth house (*mammisi*) to celebrate the king's divine birth. The whole, with service buildings, was contained by a massive mud brick wall.

The great precinct of the Temple of Karnak (the longest side, 1,837 feet [560 metres]) contains whole buildings, or parts of buildings, dating from the early 18th dynasty down to the Roman Period. Modern reconstruction work has even recovered a tiny way station of the 12th dynasty, a gem of temple building decorated with some of the finest surviving relief scenes and texts.

The Karnak complex

Of the structures on the main Karnak axis the most remarkable are the hypostyle hall and the so-called Festival Hall of Thutmose III. The former contained 134 mighty papyrus columns, 12 of which formed the higher central aisle (76 feet [23 metres]). Grill windows allowed some light to enter, but it must be supposed that even on the brightest day most of the hall was in deep gloom.

The Festival Hall is better described as a memorial hall. Its principal room is distinguished by a series of unusual columns with bell-shaped capitals, inspired by the wooden tent poles used in primitive buildings. Their lightness contrasts strikingly with the massive supports of the hypostyle hall.

Near Karnak Temple, King Akhenaton and his wife, Nefertiti, built a number of temples, later dismantled, to the sun god Aton. The vast number of blocks found in modern times indicates that these constructions were essentially open places for worship like the earlier sun temples. So, too, was the great Aton temple at Tell el-Amarna, built later in Akhenaton's reign.

The most interesting and unusual cult temple of the New Kingdom was built at Abydos by Seti I of the 19th dynasty. Principally dedicated to Osiris, it contained seven chapels dedicated to different deities, including the deified Seti himself. These chapels have well-preserved barrel ceilings and are decorated with low-relief scenes retaining much original colour.

The most remarkable monument of Ramses II, the great builder, is undoubtedly the temple of Abu Simbel (Figure 3). Although excavated from the living rock, it follows generally the plan of the usual Egyptian temple: colossal seated statues emerging from the facade, which is the cliff face; a pillared hall followed by a second leading to a vestibule; and a shrine with four statues of divinities, including one of Ramses himself.

Temple of Abu Simbel

Mention should also be made of the immense temple dedicated to the god Amon-Re at Tanis in the delta by the kings of the 21st and 22nd dynasties. Much of the stone for the so-called northern Karnak, along with colossal statues and a dozen obelisks, was appropriated from other sanctuaries in Egypt, making this a remarkable assemblage of earlier work. It was not only a cult temple but the funerary temple for the kings who were buried within the precinct.

Funerary temples. Most of the New Kingdom funerary temples were built along the desert edge in western Thebes. An exception, and by far the most original and beautiful, was Queen Hatshepsut's temple, designed and built by her steward Senenmut near the tomb of Mentuhotep II at Dayr al-Baḥrī. Three terraces lead up to the recess in the cliffs where the shrine was cut into the rock. Each terrace is fronted by colonnades of square pillars protecting reliefs of unusual subjects, including an expedition to Punt and the divine birth of Hatshepsut. Ramps lead from terrace to terrace, and the uppermost level opens into a large court with colonnades. Chapels of Hathor (the principal deity of the temple) and Anubis occupy the south and north ends of the colonnade of the second terrace.

The largest conventionally planned funerary temple was

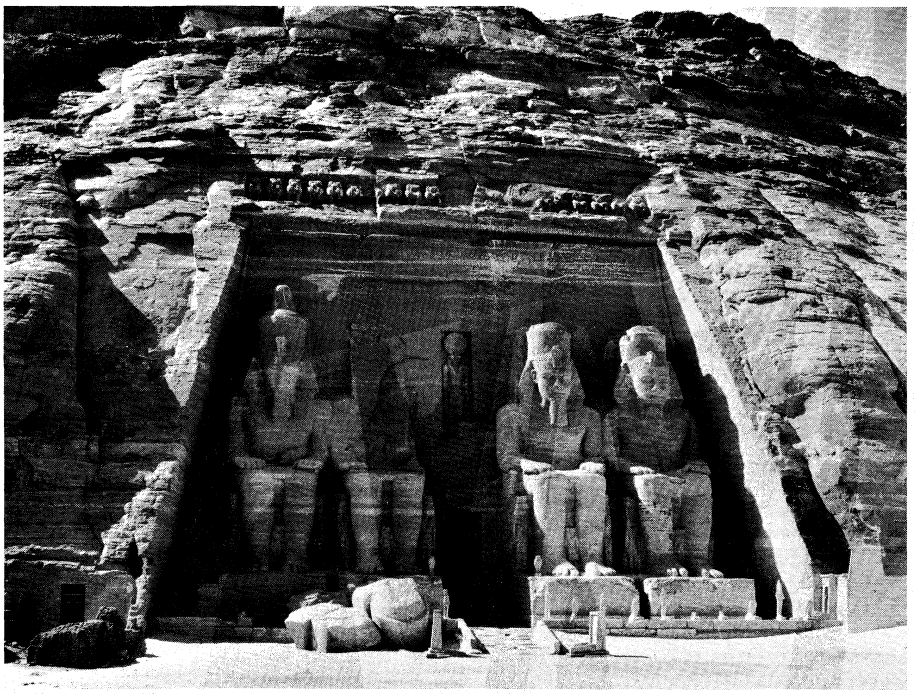


Figure 3: Entrance to the Nubian cliff temple of Ramses II at Abu Simbel, Egypt, c. 1250 BC, New Kingdom, 19th dynasty.
H. Roger-Viollet

probably that of Amenhotep III, now to be judged principally from the two huge quartzite statues, the Colossi of Memnon. These and other royal sculptures found in the ruins of the temple's courts and halls testify to the magnificence now lost. Its design, as well as much of its stone, was used by Ramses II for his own funerary temple, the Ramesseum. The huge enclosure of the latter included not only the temple but also a royal palace (only traces of which can now be seen). The temple itself contained two huge open courts, entered through towering pylons, which led to a lofty hypostyle hall and a smaller hall with astronomical carvings on the ceiling. Statues of vast size stood before the second pylon, one of which, now toppled and ruined, has been estimated as weighing more than 1,000 tons. Mud brick storerooms in the enclosure preserve ample evidence of the use of the vault in the late 2nd millennium BC.

Temple at Madīnat Habu

Ramses III's funerary temple at Madīnat Habu contains the best preserved of Theban mortuary chapels and shrines, as well as the main temple components. The most private parts of the temple, to which few had access apart from the king and his priestly representatives, begin at the sides of the first hypostyle hall, with the temple treasury and a room for the processional boat of Ramses II (a much-honoured ancestor) on the south and shrines for various deities, including Ramses III, on the north. A second pillared hall is flanked by a solar chapel and a small Osiris complex, where the king took on the personae of Re, the sun-god, and of Osiris, god of the underworld, a transfiguration considered necessary for his divine afterlife. Beyond the Osiris complex, along the temple axis, is a third small hall and the main shrine for the Theban god Amon; two lateral shrines were reserved for Amon's consort Mut and their divine child Khons.

As with most New Kingdom temples, the mural decorations on the outer walls of funerary temples, including that at Madīnat Habu, dealt mainly with the military campaigns of the king, while the inner scenes were mostly of ritual significance. Within the temple precinct lived and worked a whole community of priests and state officials. A small palace lay to the south of the main building, and a further suite of rooms for the king was installed in the castellated gate building on the east side of the precinct. The reliefs in this "high gate" suggest that the suite was used for recreational purposes by the king together with his women.

Domestic architecture. Mud brick and wood were the standard materials for houses and palaces throughout the Dynastic Period; stone was used occasionally for such architectural elements as doorjambs, lintels, column bases, and windows.

The workman's house

The best preserved private houses are those of modest size in the workmen's village of Dayr al-Madīnah. Exceptional in that they were built of stone, they typically had three or four rooms, comprising a master bedroom, a reception room, a cellar for storage, and a kitchen open to the sky; accommodation on the roof, reached by a stair, completed the plan.

Villas for important officials in Akhenaton's city of Tell el-Amarna were large and finely decorated with brightly painted murals. The house of the vizier Nakht had at least 30 rooms, including separate apartments for the master, his family, and his guests. Such houses had bathrooms and lavatories. The ceilings of large rooms were supported by painted wooden pillars, and there may have been further rooms above. Where space was restricted (as in Thebes) houses of several stories were built. Tomb scenes that show such houses also demonstrate that windows were placed high to reduce sunlight and that hooded vents on roofs were used to catch the breeze.

Palaces, as far as can be judged from remains at Thebes and Tell el-Amarna, were vast, rambling magnified versions of Nakht's villa, with broad halls, harem suites, kitchen areas, and wide courts. At Tell el-Amarna some monumental formality was introduced in the form of porticoes, colonnades, and statuary. Lavish use was made of mural and floor decoration in which floral themes predominated.

SCULPTURE

The Egyptian artist, whose skills are best exemplified in sculpture, regarded himself essentially as a craftsman. Owing to his discipline and highly developed aesthetic sense, however, the products of his craft deserve to rank as art outstanding by any standards.

Much of the surviving sculpture is funerary—statues for tombs. Most of the remainder was made for placing in temples—votive for private persons and ritual for royal and divine representations. Royal colossi were ritual and also served to proclaim the grandeur and power of the king. By itself, however, a statue could represent no one unless it carried an identification in hieroglyphs.

Types of figures

Emergence of types in the Old Kingdom. The standing male figure with left leg advanced and the seated figure were the most common types of Egyptian statuary. Traces of wooden figures found at Ṣaqqārah show that the first type was being made as early as the 1st dynasty. The earliest seated figures are two of King Khasekhem of the 2nd dynasty (Egyptian Museum, Cairo, and Ashmolean, Oxford), which, although relatively small, already embody the essential monumentality of all royal sculpture.

Supreme sculptural competence was achieved remarkably quickly. The primitive, yet immensely impressive life-size statue of Djoser (Egyptian Museum) pointed the way to the magnificent royal sculptures from the 4th-dynasty pyramid complexes at Giza. For subtlety of carving and true regal dignity scarcely anything of later date surpasses the diorite statue of Khafre (Egyptian Museum). Scarcely less fine are the sculptures of Menkaure (Mycerinus). The pair statue of the king and his wife (Museum of Fine Arts, Boston) exemplifies wonderfully both dignity and marital affection (Figure 4); the triads showing the king with goddesses and nome (provincial) deities exhibit a complete mastery of carving hard stone in many planes.

This union of skill and genius was achieved in nonroyal statuary more frequently in the Old Kingdom than later. The painted limestone statues of Prince Rahotep and his wife, Nofret (Egyptian Museum), exemplify this achievement in the formal category of seated figures. They also display the Egyptian's unsurpassed skill in inlaying eyes into sculptures, a skill further demonstrated in the wooden figure of Ka'aper, known as Shaykh al-Balad (Egyptian Museum), the very epitome of the self-important official (see Figure 8).

Scribal statues

Among additions to the sculptural repertoire during the Old Kingdom was the scribal statue. Examples in the Louvre and in the Egyptian Museum express brilliantly the alert vitality of the bureaucrat, squatting on the ground with brush poised over papyrus. The heads of such figures possess striking individuality, even if they are not true portraits.

Figure 4: King Menkaure and Queen Khamerernebty II, slate sculpture from Giza, Egypt, *c.* 2525 BC, Old Kingdom, 4th dynasty. In the Museum of Fine Arts, Boston. Height 1.42 m.
By courtesy of the Museum of Fine Arts, Boston. Museum Expedition

Refinements of the Middle Kingdom. Changes in funerary practices during the Middle Kingdom led to a reduction in the number of sculptures. Royal sculptures, particularly of Sesostris III and Amenemhet III (British Museum), achieved a high degree of realism, even of portraiture. The first true royal colossi were produced in the 12th dynasty (if the Great Sphinx of Giza is discounted) for the embellishment of cult temples. Colossi of Amenemhet I and Sesostris I (Egyptian Museum) exhibit a hard, uncompromising style said to typify the ruthless drive of the 12th-dynasty kings.

In this period, too, the sphinx—the recumbent lion with head or face of the king—became a commonly used image of the king as protector. The great red granite sphinx of Amenemhet II from Tanis (Louvre) expresses the idea most potently.

In private sculpture during the Middle Kingdom the subject is usually portrayed seated or squatting, occasionally standing, and wearing an all-enveloping cloak. The body was mostly concealed, but its contours were often subtly suggested in the carving, as in the figure of Khertyhotep (Ägyptisches Museum, West Berlin). Of female subjects, none is more impressive than that of Sennu (Museum of Fine Arts, Boston), a wonderful example of a figure in repose.

Block statues

The simplification of the human figure was carried to its ultimate in the block statue, a uniquely Egyptian type that represents the subject squatting on the ground with knees drawn up close to his body. The arms and legs may be wholly contained within the cubic form, hands and feet alone discretely protruding. The 12th-dynasty block statue of Sihathor (British Museum) is the earliest dated example.

Innovation, decline, and revival in the New Kingdom. Excellence of craftsmanship is the hallmark of 18th-dynasty sculpture, in a revival of the best traditions of the Middle Kingdom. Wonderfully sensitive statues of Hatshepsut and Thutmose III confirm the return of conditions in which great work can be achieved. A seated limestone statue of Hatshepsut (Metropolitan Museum of Art, New York City) shows the queen as king, but with an expression of consummate grace. A schist statue of Thutmose III (Luxor Museum), in the perfection of its execution and subtlety of its realization, epitomizes regality.

The placing of votive statues in temples led to a proliferation of private sculptures during the New Kingdom. The sculptures of Senenmut, steward of Hatshepsut, exemplify the development. At least 23 votive statues (some fragmentary) of this royal favourite are known, exhibiting many different forms.

Colossal sculpture

Colossal sculpture, which reached its apogee in the reign of Ramses II, was used to splendid, and perhaps less bombastic, effect by Amenhotep III. The great sculptures of his funerary temple, already mentioned, including the immense Colossi of Memnon, were part of the noble designs of his master of works, also called Amenhotep (son of Hapu). Most unusually, this distinguished commoner was allowed a funerary temple for himself and larger-than-life votive sculptures (Egyptian Museum and Luxor Museum) that show him in contrasting attitudes, as stern-faced authoritarian and as submissive scribe.

The realistic portraiture that can be noted in certain sculptures of Amenhotep III (British Museum) hints of an artistic change that was developed in the subsequent reign of Akhenaton. The distinctive style of this period has come to be called Amarna, after the location of Akhenaton's new capital in Middle Egypt. Colossal sculptures of the King from the dismantled Karnak temples (Egyptian Museum) emphasize his bodily peculiarities—elongated facial features, heavy breasts, and swelling hips (Figure 5). Sculptures of Nefertiti, his queen, are often executed in the most remarkably sensual manner (*e.g.*, the Louvre torso). Sculptures from later in the reign display innovations of style with no loss of artistry, at the same time avoiding the grotesqueries of the early years. Of this period is the famous painted bust of Nefertiti (Ägyptisches Museum).

Much of the best of the artistic legacy of Akhenaton's reign persisted in the sculpture of subsequent reigns—Tutankhamen, Horemheb, and the early kings of the 19th dynasty—but a marked change came in the reign of Ramses II. It is a commonplace to decry the quality of his monumental statuary, although little in Egypt is more dra-

The reign of Ramses II

Figure 5: King Akhenaton, sandstone pillar statue from the Temple of Aton at Karnak, Egypt, New Kingdom, 18th dynasty (mid-14th century BC). In the Egyptian Museum, Cairo. Height 4.00 m.
Hirmer Fotoarchiv, Munchen

matic and compelling than the great seated figures of this king at Abu Simbel. Nevertheless, there is much truth in the belief that the steady decline in sculpture began during Ramses II's reign. Royal portraiture subsequently became conventional. Occasionally a sculptor might produce some unusual piece, such as the extraordinary figure of Ramses VI with his lion, dragging beside him a Libyan prisoner (Egyptian Museum). Among private sculptures there is the scribal statue of Ramsesnakht (Egyptian Museum); the subject bends over his papyrus while Thoth (the divine scribe), in baboon form, squats behind his head.

A change was to come with the advent of the Kushite (Nubian) kings of the 25th dynasty. The portraiture of the Kushite kings exhibits a brutal realism that may owe much to the royal sculpture of the 12th dynasty; the sphinx of Taharqa, fourth king of the 25th dynasty (British Museum), is a good example.

Archaism is strikingly evident in the private sculpture of the last dynasties. Types of statue common in the Middle Kingdom and 18th dynasty were revived, and many very fine pieces were produced. The sculptures of the mayor of Thebes, Montemhat (Egyptian Museum), display great variety, excellent workmanship, and, in one case, a realism that transcends the dictates of convention.

In considering the clear sculptural qualities of Late Period work one should never overlook the primary purpose of most Egyptian sculpture: to represent the individual in death before Osiris, or in life and death before the deities of the great temples. To this end the statue was not only a physical representation but also a vehicle for appropriate texts, which might be inscribed obtrusively over beautifully carved surfaces. The extreme example of such "disfigurement" is a so-called healing statue (Louvre) of which even the wig is covered with texts.

RELIEF SCULPTURE AND PAINTING

For Egyptians the decoration of tomb walls with reliefs or painted scenes provided some certainty of the perpetuation of life; in a temple, similarly, it was believed that mural decoration magically ensured the performance of important ceremonies and reinforced the memory of royal deeds.

Importance of mural decoration

The beginnings of the dynastic tradition can be found in tombs of the 3rd dynasty, such as that of Hesire at Ṣaqqārah; it contained mural paintings of funerary equipment and wooden panels carrying figures of Hesire in the finest low relief (Egyptian Museum; see Figure 8). Generally speaking, mural decorations were in paint when the ground was mud brick or stone of poor quality, and in relief when the walls were in good stone. Painting and drawing formed the basis of what was to be carved in relief, and the finished carving was itself commonly painted.

In tombs the mural decorations might be left unfinished, being only partly sketched or partly carved by the time of the burial. Uncompleted scenes reveal clearly the methods of laying out walls for decoration. The prepared wall was marked out with red guidelines, the grid described earlier being used for major human figures and sometimes for minor ones. Preliminary outlines were corrected and paint was applied usually in tempera, pigments being mostly mineral-based.

In the Old Kingdom pure painting of the highest quality is found as early as the 4th dynasty in the scene of geese from the tomb of Nefermaat and Atet at Maydum. But the glory of Old Kingdom mural decoration is the low-relief work in the royal funerary monuments of the 5th dynasty and in the private tombs of the 5th and 6th dynasties in the Memphite necropolis. Outstanding are the reliefs from the sun temple of King Neuserre at Abu Jīrab (Ägyptisches Museum, East and West Berlin) and the scenes of daily life in the tombs of Ptahhotep and Ti at Ṣaqqārah.

The tradition of fine painting was continued in the Middle Kingdom. At Beni Hasan the funerary chambers are crowded with paintings exhibiting fine draftsmanship and use of colour. The best relief work of the period, reviving the Memphite tradition, is found at Thebes in the tomb of Mentuhotep II at Dayr al-Baḥrī and in the little shrine of Sesostris I at Karnak, where the fine carving is greatly enhanced by a masterly use of space in the disposition of figures and text.

In the early 18th dynasty the relief tradition was revived at Thebes and can best be observed in the carvings in Hatshepsut's temple at Dayr al-Baḥrī. Later royal reliefs of Amenhotep III and of the post-Amarna kings show a stylistic refinement that was carried to its best in the reign of Seti I, at Karnak, at Abydos, and in his tomb at Thebes.

The 18th dynasty also saw Egyptian painting reach its highest achievement in the tombs of the nobles at Thebes (Figure 6). The medium of decoration and an apparently

High point of Egyptian painting

By courtesy of the Trustees of the British Museum

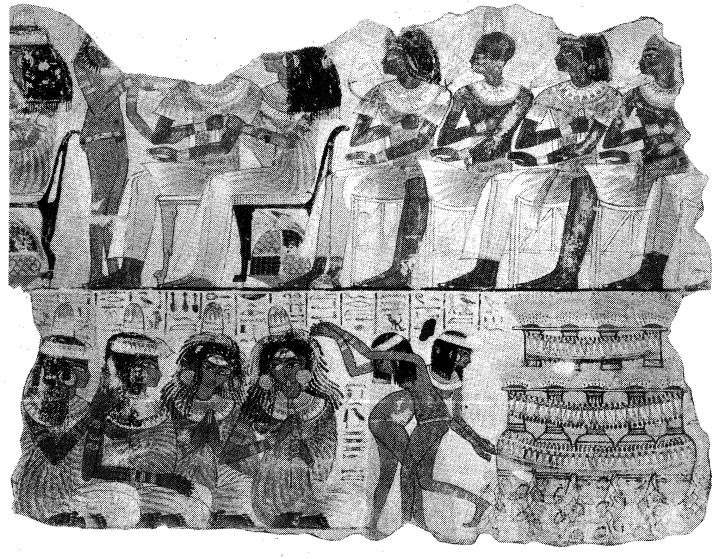


Figure 6: Banquet scene with musicians, tempera painting on gesso from the tomb of Nebamun at Thebes, 18th dynasty (*c.* 1400 BC). In the British Museum.

greater artistic freedom led to the introduction of small, often entertaining details into standard scenes. The tiny tombs of Menna and Nakht are full of such playful vignettes. The paintings in great tombs, such as that of Rekhmire, are more formal but still crammed with unusual detail. Fragments of mural and floor paintings from palaces and houses at Thebes and Tell el-Amarna provide tantalizing glimpses of the marsh and garden settings of everyday upper-class life.

The fine royal reliefs of the late 18th dynasty were matched by those in private tombs at Thebes (Ramose and Kheruef) and Ṣaqqārah (Horemheb); these are breathtaking in execution and, in the case of Horemheb, both moving and original. Interest in relief subsequently passed to the work in the temples of the 19th and 20th dynasties. The most dramatic subject was war, whether the so-called triumph of Ramses II at Kadesh (Thebes and Abu Simbel), or the more genuine successes of Ramses III against the Libyans and the Sea Peoples (Madīnat Habu). The size and vitality of these ostentatious scenes are stupendous, even if their execution tends to be slapdash.

The artistic renaissance of the 25th and 26th dynasties is less evident in painting and relief than in sculpture. Although the fine work in the tomb of Montemhat at Thebes is distinctly archaizing, it is, nevertheless, exceptional in quality. The skills of the Egyptian draftsman, nurtured by centuries of exercise at large and small scale, remained highly professional. This skill is seen at its most consistent level in the illumination of papyruses. The practice of including drawings, often painted, in religious papyruses flourished from the time of the 18th dynasty and reached a high point around 1300 BC. The peak of achievement is probably represented by the *Book of the Dead* of the scribe Ani (British Museum), in the vignettes of which both technique and the use of colour are outstanding. Subsequently, and especially in the Late Period, pure line drawing was increasingly employed.

Illumination of papyruses

PLASTIC ARTS

In Egypt pottery provided the basic material for vessels of all kinds. Fine wares and many other small objects were made from faience. Glass arrived late on the scene and was used somewhat irregularly from the New Kingdom onward.

Pottery. Generally speaking, Egyptian pottery had few artistic pretensions. In the tomb of Tutankhamen most of the pottery vessels were simple wine jars in the form of amphorae. It is surprising that no finer pottery vessels were found, because high-quality ware was made during the late 18th and 19th dynasties, often brightly painted with floral designs.

Pottery was rarely modeled, although human and animal figures occur in small numbers throughout the Dynastic Period. Small vessels in animal form were also made, especially during the Middle and New Kingdoms, and a fine category of highly burnished red pottery vases in female form was produced during the 18th dynasty.

Faience. The place of pottery for modeling was filled with faience (a glazed composition of ground quartz), most commonly blue or green in colour. In the Early Dynastic Period it was much used for the making of small animal and human figures, and throughout the Dynastic Period it continued to be used in this way, among the most striking results being the blue-glazed hippopotamus figures of Middle Kingdom date.

Hippopotamus figures

In the Late Period, in particular, the making of amulets and divine figurines in faience was highly developed, and many pieces display a high standard of modeling and perfection of glazing. The vast quantities of *ushabti* (*shabti*, or *shawabti*) figures provided as parts of funerary equipments are mostly routine work, but the finest examples from the New Kingdom, and some of Saite date, show complete mastery of a difficult technique.

Faience tiles were also first made in the early dynasties and were used chiefly for wall decoration, as in the subterranean chambers of the Step Pyramid. In the New Kingdom, tiles with floral designs were used in houses and palaces in the reigns of Amenhotep III and his successors. During the 19th and 20th dynasties royal palaces at Per Ramessu (modern Qantīr), Tell al-Yahudīyah, and Madīnat Habu were embellished with remarkable polychrome tiles, many of which bear figures of captive foreigners.

Throughout the Dynastic Period faience was regularly used for simple beads, amulets, and other components of jewelry. Quite exceptional is the extraordinary *was*-sceptre (a symbol of divine power) found at Tūkh, near Naqādah (Victoria and Albert Museum, London). It is dated to the reign of Amenhotep II and originally measured about six and a half feet (two metres) in length.

Glass. In the form of glaze, glass was known to the ancient Egyptians from early predynastic times, but the material was not used independently until the 18th dynasty. From the mid-18th dynasty and during the 19th dynasty glass was used for small amulets, beads, inlays, and especially for small vessels. The material was opaque, blue being the predominant colour, although other bright colours were also achieved. The vessels, made around sand cores, were mostly drinking cups or flasks for precious liquids and were often decorated with trailed patterns applied as glass threads. Glass was certainly a material of luxury, a fact confirmed by the presence of two glass goblets with gold rims among a treasure of precious vessels from the reign of Thutmose III (Metropolitan Museum of Art).

The use of glass for inlay is notably demonstrated in Tutankhamen's golden throne, in his solid gold mask, and in much of his jewelry (Figure 7). After the 19th dynasty, glass manufacture seems largely to have been discontinued until the Late Period,when the use of glass for inlays was revived.

Glass inlay

DECORATIVE ARTS

Jewelry. Gold provided Egyptian jewelry with its richness; it was used for settings, cloisonné work, chains, and beads, both solid and hollow. Soldering, granulation, and wire making were practiced. Precious stones were not used, but a wide range of semiprecious stones was exploited: carnelian, amethyst, garnet, red and yellow jasper, lapis lazuli, feldspar, turquoise, agate. Additional colours and textures were provided by faience and glass.

Ancient Egyptian jewelers had a fine eye for colour and an excellent sense of design. From the earliest dynasties come bracelets from the tomb of King Djer at Abydos (Egyptian Museum); from the 4th dynasty, the armlets of Queen Hetepheres, of silver inlaid with carnelian, turquoise, and lapis lazuli (Egyptian Museum and Museum of Fine Arts). There are examples of splendid and delicate jewelry dating from the Middle Kingdom: in particular, pieces found at Dahshūr and al-Lāhūn—circlets of Princess Khnumet

By courtesy of the Egyptian Museum, Cairo.
Photograph, Eliot Elisofon, *Life*, c. 1948, Time, Inc.

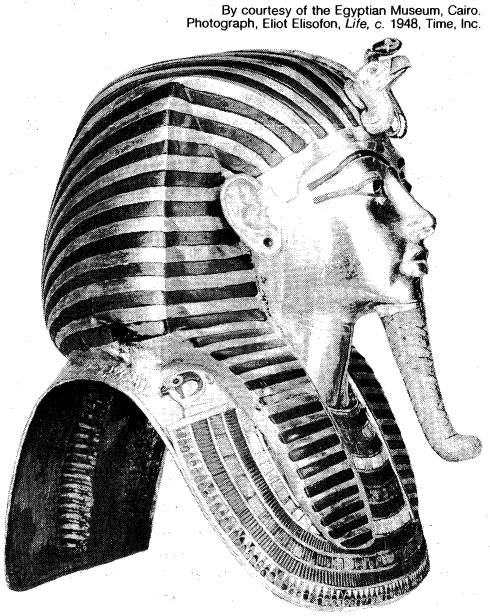


Figure 7: Gold funerary mask of the pharaoh Tutankhamen inlaid with lapis lazuli and coloured glass, New Kingdom, 18th dynasty (c. 1323 BC). In the Egyptian Museum, Cairo. Height 53.3 cm.

(Egyptian Museum), pectorals of Princess Sithathor and Queen Meret (Egyptian Museum), and girdles of Princess Sithathor-iunet (Metropolitan Museum of Art).

The large and spectacular collection of jewelry buried with Queen Ahhotep of the early 18th dynasty (Egyptian Museum) includes many unusual designs; her gold chain is a masterpiece. Much fine 18th-dynasty jewelry has survived, but all is dominated by that of Tutankhamen (Egyptian Museum). This huge collection demonstrates all the techniques of the goldsmith's and the lapidary's arts.

Copper and bronze. The techniques of metalworking were probably introduced into Egypt from the Middle East at a very early date. At first copper was most commonly used; but from at least the late 3rd millennium it was often alloyed with tin, as bronze.

The skill and artistry of the metalworker is shown in the fine bowls, jugs, and other vessels from all periods, and in statues and statuettes of gods, kings, and ordinary mortals. Most vessels were made by raising from metal ingots, beaten on wooden anvils. In the Late Period many vessels were produced by casting. Huge situlae, vessels used for carrying sacred liquids, are often decorated with scenes and inscriptions.

The earliest and largest metal figure from Egypt is the life-size statue of Pepi I (Egyptian Museum) made of copper plates fitted to a wooden core, the plates probably beaten, not cast. Casting in open molds was developed early for tools and weapons, but the lost-wax process (*cire-perdue*), using closed molds, was not employed until the Middle Kingdom. Even in the 18th dynasty the casting of bronze figures occurred on a relatively small scale.

Casting of figures

The casting of large-scale bronze figures achieved its highest point in the late New Kingdom down to the 25th dynasty. The outstanding example from this period is the figure of Karomama (Louvre). The exceptionally elegant modeling of the female form is greatly enriched by inlays of gold and silver reproducing the feathered pattern of the gown and an elaborate collar of floral motifs.

In the Late Period huge numbers of excellent castings of conventional sacred figures and animals were produced. The so-called Gayer-Anderson cat (British Museum) is technically and artistically without peer.

Gold and silver. Gold was more easily obtainable in ancient Egypt than silver and was therefore less valuable (until the late New Kingdom). Gold was also easier to work and unaffected by environmental conditions. In consequence, many more gold than silver objects have survived.

Decorative uses of gold

Apart from jewelry, gold was lavishly used for many decorative purposes, as thin sheet, leaf, and inlay, in funerary equipment, and for vessels and furniture. The range of uses is best exemplified in the objects from the tomb of Tutankhamen.

The gold-plated, gold-inlaid furniture of Queen Hetepheres of 4th-dynasty date reveals how early Egyptian craftsmen mastered the working of gold. Gold vessels have rarely survived, but those from the royal burials of Tanis (Egyptian Museum) preserve styles and techniques that go back to the traditions of the New Kingdom and earlier. Gold statuettes also are rare, but again, surviving examples, such as the magnificent falcon head of a cult statue of 6th-dynasty date from Hierakonpolis and the divine triad of Osiris, Isis, and Horus of the 22nd dynasty (Louvre), show the achievements of early and late times.

In a hoard of precious vessels found at Bubastis and dated to the 19th dynasty (Egyptian Museum) there were three silver pieces of exceptional interest, in particular a jug the handle of which is of gold and in the shape of a goat. Greater availability of silver in later times is demonstrated by two massive silver coffins and a number of vessels in the royal burials at Tanis (Egyptian Museum).

Wood. The wooden sculpture of the Old Kingdom shows the carver of wood at his most skillful and sensitive (Figure 8). But it is in the field of cabinetmaking that the ancient woodworker excelled. Best known are the many chairs, tables, stools, beds, and chests found in Tutankhamen's tomb. Many of the designs are exceptionally practical and elegant. Techniques of inlay, veneering, and marquetry are completely mastered. One chest is veneered with strips of ivory and inlaid with 33,000 small pieces

Cabinetmaking

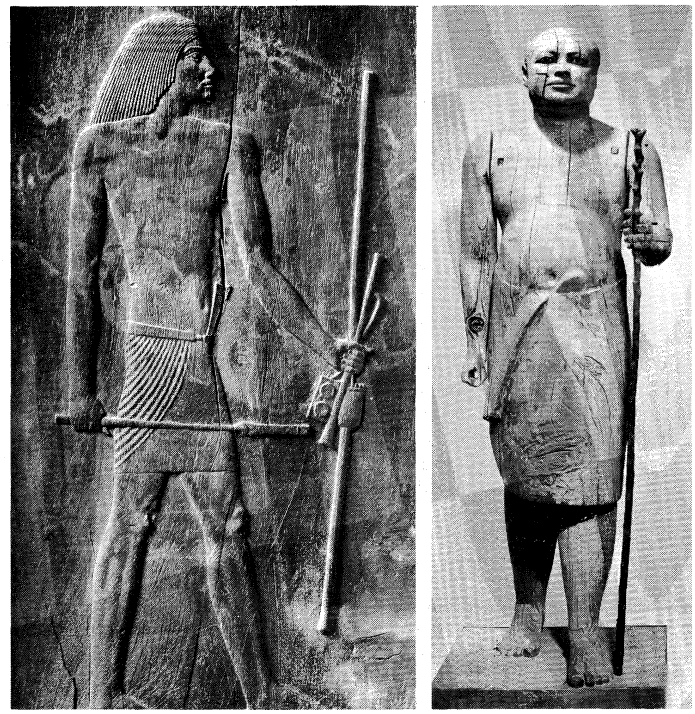


Figure 8: (Left) Wood relief of the scribe Hesire, from the Tomb of Hesire at Ṣaqqārah, 3rd dynasty. In the Egyptian Museum, Cairo. Height 1.14 m. (Right) Shaykh al-Balad, wood statue from Ṣaqqārah, 5th dynasty. In the Egyptian Museum. Height 1.10 m.
Photographs, Hirmer Fotoarchiv, Munchen

of ivory and ebony. Fine furniture was being produced in very early times, as is confirmed by the skillfully restored furniture from the secondary burial of Hetepheres (Egyptian Museum and Museum of Fine Arts).

Among the most charming and delicate products of the Egyptian woodworker are the many toilet spoons and containers in the form of graceful swimming girls, lute players in the marshes, and fishes and animals. At the other extreme, nothing is more remarkable than the great boat, more than 140 feet (43 metres) long, found in a trench by the side of the Great Pyramid.

Ivory and bone. Of the few small ivory figurines to have survived from pharaonic times, two royal representations found in the Early Dynastic temple at Abydos (Egyptian Museum and British Museum) are outstanding. There can be little doubt, in spite of the paucity of survivals, that fine decorative objects of ivory were made at all periods. A gazelle and a grasshopper of the 18th dynasty (Metropolitan Museum of Art and Brooklyn Museum) may truly be described as *objets de vertu.* Many fine examples of the use of ivory were found in Tutankhamen's tomb, from simple geometric marquetry patterns to box panels carved with exquisitely informal scenes of the king with his queen.

Ivory figurines

Greco-Roman Egypt

After the conquest of Egypt by Alexander the Great, the independent rule of Pharaohs in the strict sense came to an end. Under the Ptolemies, whose rule followed Alexander's, profound changes took place in art and architecture.

The most lasting impression of the new period is made by the architectural legacy. Although very little survives of important funerary architecture, there is a group of tombs at Tunah al-Jabal of unusual form and great importance. Most interesting is the tomb of Petosiris, high priest of Thoth in nearby Hermopolis Magna in the late 4th century BC. It is in the form of a small temple with pillared portico, elaborate column capitals, and a large forecourt. In its mural decorations a strong Greek influence merges with the traditional Egyptian modes of expression.

A boom in temple building of a more conventional kind followed the establishment of the Ptolemaic regime. At Dandarah, Esna, Idfū, Kawm Umbū (Kôm Ombo), and

Ptolemaic temples

Philae the Egyptian cult temple can be studied better than at almost any earlier temple. The temple of Horus at Idfū is the most complete, displaying all the essentials of the classical Egyptian temple. The common judgment on these late temples is adverse, but for exploitation of setting and richness of detail it is difficult to fault the temples of Philae and Kawm Umbū, in particular.

In relief carving a noticeable change had taken place in the conventional proportions of human figures during the Saite Period, and subsequently, with added influences from Greek art, a more voluptuous style of human representation developed. There is also undoubtedly some coarseness in much of the new work. Nevertheless, there is much to admire in the best reliefs of the Hathor Temple at Dandarah and in the double cult temple of Sebek and Horus at Kawm Umbū.

Generous representation of the human form, especially the female form, also characterizes the sculpture of the Ptolemaic Period, and there is little to match the figure of Queen Arsinoe II (State Hermitage Museum, Leningrad). It is in the treatment of the head, however, that the greatest changes took place. It is a matter of debate whether the new emphasis on portraiture was attributable to influences from the classical world or was a development of earlier Egyptian sculptural tendencies. Fine pieces such as the schist "green" head of a man (Ägyptisches Museum) could not have failed to impress the observer from the Ptolemaic court or the later Roman administration. One of the finest surviving heads, in diorite, slightly larger than life-size and of dominating appearance, is the "black" head in The Brooklyn Museum (Figure 9).

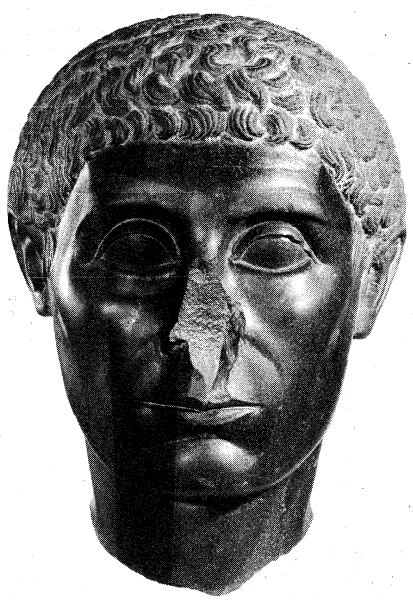


Figure 9: Black diorite head of a high official, possibly a high priest of Ptah of Memphis, Ptolemaic Period (c. 75 BC). In The Brooklyn Museum. Height 41.4 cm.

By courtesy of The Brooklyn Museum, gift of the Charles Edwin Wilbour Fund

Throughout the Ptolemaic Period votive sculpture of private persons was made in great quantity. After the Roman conquest it became rare and of indifferent quality. Such Egyptian art as can be isolated in the Roman Period is found in funerary equipment—in coffins, shrouds, and panel portraits. A mixture of Egyptian and classical styles and of diverse symbolisms can be observed. The great shroud showing the deceased and his mummy protected by the mortuary deity, Anubis (Louvre), while harking back to the traditions of pharaonic Egypt, also displays in the figure of the deceased a style that points to Byzantium.

Mummy portraits

The mummy, or Fayum, portraits are Egyptian only in that they are associated with essentially Egyptian burial customs. Painted in an encaustic technique, they represent mostly Greek inhabitants of Egypt. Seen properly in context, as in the complete mummy of Artemidorus (British Museum), they provide a strange epilogue to the funerary art of 3,000 years of pharaonic Egypt. In this field and in a few others the vigour of the native tradition persisted artistically up to the Roman conquest. Thereafter the decline was rapid and complete. By the 3rd century AD Egypt was on the way to becoming a Christian country. The old tradition was not only destroyed, it was no longer valued. Coptic art was to find its inspiration elsewhere.

BIBLIOGRAPHY. The best general surveys are CYRIL ALDRED, *Egyptian Art, in the Days of the Pharaohs, 3100–320 B.C.* (1980); CYRIL ALDRED (*et al.*), *Le Temps des pyramides: de la préhistoire aux Hyksos, 1560 av. J.-C.* (1978), *L'Empire des conquérants: l'Égypte au Nouvel Empire (1560–1070)* (1979), and *L'Égypte du crépuscule: de Tanis à Méroé, 1070 av. J.-C.–IVe siècle apr. J.-C.* (1980); KAZIMIERZ MICHALOWSKI, *The Art of Ancient Egypt,* trans. and adapted from the Polish and French (1969); KURT LANGE and MAX HIRMER, *Egypt: Architecture, Sculpture, Painting in Three Thousand Years,* 4th ed. (1968; originally published in German, 4th ed., 1967); WILLIAM STEVENSON SMITH, *The Art and Architecture of Ancient Egypt,* rev. ed. by WILLIAM KELLY SIMPSON (1983); CLAUDE VANDERSLEYEN, *Das alte Ägypten* (1975); and WALTHER WOLF, *Die Kunst Aegyptens* (1957). On conventions and general principles, fundamental works are ERIK IVERSEN, *Canon and Proportions in Egyptian Art,* 2nd ed. rev. (1975); HEINRICH SCHÄFER, *Principles of Egyptian Art,* ed. by EMMA BRUNNER-TRAUT (1974; originally published in German, 4th ed., 1963); and WILLIAM STEVENSON SMITH, *Interconnections in the Ancient Near-East: A Study of the Relationships Between the Arts of Egypt, the Aegean, and Western Asia* (1965). The only comprehensive work on architecture is ALEXANDER BADAWY, *A History of Egyptian Architecture,* 3 vol. (1954–68); for a thoughtful study, see E. BALDWIN SMITH, *Egyptian Architecture as Cultural Expression* (1938, reissued 1968); on the pyramids, in particular, the best introduction is I.E.S. EDWARDS, *The Pyramids of Egypt,* rev. ed. (1986). For the best introduction to the analysis of sculpture, see HANS GERHARD EVERS, *Staat aus dem Stein: Denkmäler, Geschichte und Bedeutung der ägyptischen Plastik während des Mitteleren Reichs,* 2 vol. (1929). For Old Kingdom to New Kingdom sculpture, see CYRIL ALDRED, *Old Kingdom Art in Ancient Egypt* (1949), *Middle Kingdom Art in Ancient Egypt, 2300–1590 B.C.* (1950), and *New Kingdom Art in Ancient Egypt During the Eighteenth Dynasty: 1590 to 1315 B.C.* (1951), varying editions reissued as *The Development of Ancient Egyptian Art, from 3200 to 1315 B.C.,* 3 vol. in 1 (1952, reprinted 1973); and JACQUES VANDIER, *Manuel d'archéologie égyptienne,* vol. 3, *Les Grandes Époques: la statuaire* (1958). For the Old Kingdom, see WILLIAM STEVENSON SMITH, *A History of Egyptian Sculpture and Painting in the Old Kingdom* (1946, reissued 1978); and for the Late Period and Greco-Roman Period, BERNARD V. BOTHMER (comp.), *Egyptian Sculpture of the Late Period, 700 B.C. to A.D. 100* (1969). Excellent reproductions of paintings and drawings are to be found in NINA M. DAVIES and ALAN H. GARDINER, *Ancient Egyptian Paintings,* 3 vol. (1936); good surveys and some unusual material are in EMMA BRUNNER-TRAUT, *Egyptian Artists' Sketches* (1979). Also see T.G.H. JAMES, *Egyptian Painting* (1985); ARPAG MEKHITARIAN, *Egyptian Painting* (1954, reissued 1978; originally published in French, 1954); and WILLIAM H. PECK, *Drawings from Ancient Egypt* (1978). A good survey of the whole range of pottery is JANINE BOURRIAU, *Umm el-Ga'ab: Pottery from the Nile Valley Before the Arab Conquest* (1981). On glassware, see JOHN D. COONEY, *Glass* (1976). Jewelry is well treated artistically and technically in CYRIL ALDRED, *Jewels of the Pharaohs* (1971, reissued 1978); and technically and archaeologically in ALIX WILKINSON, *Ancient Egyptian Jewellery* (1971, reissued 1975). An excellent general account of furniture is contained in HOLLIS S. BAKER, *Furniture in the Ancient World: Origins and Evolution 3100–475 B.C.* (1966); for a reliable technical study, see G. KILLEN, *Ancient Egyptian Furniture,* vol. 1 (1980). On Greco-Roman art there is an excellent summary in GÜNTHER GRIMM, *Kunst der Ptolemäer- und Römerzeit im Ägyptischen Museum Kairo* (1975); for useful background essays, see HERWIG MAEHLER and VOLKER MICHAEL STROCKA (eds.), *Das ptolemäische Ägypten* (1978); and on reliefs in Greco-Roman temples, see ERICH WINTER, *Untersuchungen zu den ägyptischen Tempelreliefs der griechisch-römischen Zeit* (1968). The essential work on Fayum portraits is KLAUS PARLASCA, *Mumienporträts und verwandte Denkmäler* (1966).

(T.G.H.J.)

Einstein

Recognized in his own time as one of the most creative intellects in human history, Albert Einstein, in the first 15 years of the 20th century, advanced a series of theories that for the first time asserted the equivalence of mass and energy and proposed entirely new ways of thinking about space, time, and gravitation. His theories of relativity and gravitation were a profound advance over the old Newtonian physics and revolutionized scientific and philosophic inquiry.

Herein lay the unique drama of Einstein's life. He was a self-confessed lone traveller; his mind and heart soared with the cosmos, yet he could not armour himself against the intrusion of the often horrendous events of the human community. Almost reluctantly he admitted that he had a "passionate sense of social justice and social responsibility." His celebrity gave him an influential voice that he used to champion such causes as pacifism, liberalism, and Zionism. The irony for this idealistic man was that his famous postulation of an energy–mass equation, which states that a particle of matter can be converted into an enormous quantity of energy, had its spectacular proof in the creation of the atomic and hydrogen bombs, the most destructive weapons ever known.

Einstein.

By courtesy of the Nobelstiftelsen, Stockholm

Early life and career. Albert Einstein was born in Ulm, Germany, on March 14, 1879. The following year his family moved to Munich, where Hermann Einstein, his father, and Jakob Einstein, his uncle, set up a small electrical plant and engineering works. In Munich Einstein attended rigidly disciplined schools. Under the harsh and pedantic regimentation of 19th-century German education, which he found intimidating and boring, he showed little scholastic ability. At the behest of his mother, Einstein also studied music; though throughout life he played exclusively for relaxation, he became an accomplished violinist. It was then only Uncle Jakob who stimulated in Einstein a fascination for mathematics and Uncle Cäsar Koch who stimulated a consuming curiosity about science.

By the age of 12 Einstein had decided to devote himself to solving the riddle of the "huge world." Three years later, with poor grades in history, geography, and languages, he left school with no diploma and went to Milan to rejoin his family, who had recently moved there from Germany because of his father's business setbacks. Albert Einstein resumed his education in Switzerland, culminating in four years of physics and mathematics at the renowned Federal Polytechnic Academy in Zürich.

After his graduation in the spring of 1900, he became a Swiss citizen, worked for two months as a mathematics teacher, and then was employed as examiner at the Swiss patent office in Bern. With his newfound security, Einstein married his university sweetheart, Mileva Marić, in 1903.

First contributions to science

Early in 1905 Einstein published in the prestigious German physics monthly *Annalen der Physik* a thesis, "A New Determination of Molecular Dimensions," that won him a Ph.D. from the University of Zürich. Four more important papers appeared in *Annalen* that year and forever changed man's view of the universe.

The first of these, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen" ("On the Motion—Required by the Molecular Kinetic Theory of Heat—of Small Particles Suspended in a Stationary Liquid"), provided a theoretical explanation of Brownian motion. In "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt" ("On a Heuristic Viewpoint Concerning the Production and Transformation of Light"), Einstein postulated that light is composed of individual quanta (later called photons) that, in addition to wavelike behaviour, demonstrate certain properties unique to particles. In a single stroke he thus revolutionized the theory of light and provided an explanation for, among other phenomena, the emission of electrons from some solids when struck by light, called the photoelectric effect.

Einstein's special theory of relativity, first printed in "Zur Elektrodynamik bewegter Körper" ("On the Electrodynamics of Moving Bodies"), had its beginnings in an essay Einstein wrote at age 16. The precise influence of work by other physicists on Einstein's special theory is still controversial. The theory held that, if, for all frames of reference, the speed of light is constant and if all natural laws are the same, then both time and motion are found to be relative to the observer.

In the mathematical progression of the theory, Einstein published his fourth paper, "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?" ("Does the Inertia of a Body Depend Upon Its Energy Content?"). This mathematical footnote to the special theory of relativity established the equivalence of mass and energy, according to which the energy E of a quantity of matter, with mass m, is equal to the product of the mass and the square of the velocity of light, c. This relationship is commonly expressed in the form $E = mc^2$.

Public understanding of this new theory and acclaim for its creator were still many years off, but Einstein had won a place among Europe's most eminent physicists, who increasingly sought his counsel, as he did theirs. While Einstein continued to develop his theory, attempting now to encompass with it the phenomenon of gravitation, he left the patent office and returned to teaching—first in Switzerland, briefly at the German University in Prague, where he was awarded a full professorship, and then, in the winter of 1912, back at the Polytechnic in Zürich. He was later remembered from this time as a very happy man, content in his marriage and delighted with his two young sons, Hans Albert and Edward.

In April 1914 the family moved to Berlin, where Einstein had accepted a position with the Prussian Academy of Sciences, an arrangement that permitted him to continue his researches with only the occasional diversion of lecturing at the University of Berlin. His wife and two sons vacationed in Switzerland that summer and, with the eruption of World War I, were unable to return to Berlin. A few years later this enforced separation was to lead to divorce. Einstein abhorred the war and was an outspoken critic of German militarism among the generally acquiescent academic community in Berlin, but he

was primarily engrossed in perfecting his general theory of relativity, which he published in *Annalen der Physik* as "Die Grundlagen der allgemeinen Relativitätstheorie" ("The Foundation of the General Theory of Relativity") in 1916. The heart of this postulate was that gravitation is not a force, as Newton had said, but a curved field in the space–time continuum, created by the presence of mass. This notion could be proved or disproved, he suggested, by measuring the deflection of starlight as it travelled close by the Sun, the starlight being visible only during a total eclipse. Einstein predicted twice the light deflection that would be accountable under Newton's laws.

General theory of relativity

His new equations also explained for the first time the puzzling irregularity—that is, the slight advance—in the planet Mercury's perihelion, and they demonstrated why stars in a strong gravitational field emitted light closer to the red end of the spectrum than those in a weaker field.

While Einstein awaited the end of the war and the opportunity for his theory to be tested under eclipse conditions, he became more and more committed to pacifism, even to the extent of distributing pacifist literature to sympathizers in Berlin. His attitudes were greatly influenced by the French pacifist and author Romain Rolland, whom he met on a wartime visit to Switzerland. Rolland's diary later provided the best glimpse of Einstein's physical appearance as he reached his middle 30s:

> Einstein is still a young man, not very tall, with a wide and long face, and a great mane of crispy, frizzled and very black hair, sprinkled with gray and rising high from a lofty brow. His nose is fleshy and prominent, his mouth small, his lips full, his cheeks plump, his chin rounded. He wears a small cropped mustache. (By permission of Madame Marie Romain Rolland.)

Einstein's view of humanity during the war period appears in a letter to his friend, the Austrian-born Dutch physicist Paul Ehrenfest:

> The ancient Jehovah is still abroad. Alas, he slays the innocent along with the guilty, whom he strikes so fearsomely blind that they can feel no sense of guilt. . . . We are dealing with an epidemic delusion which, having caused infinite suffering, will one day vanish and become a monstrous and incomprehensible source of wonderment to later generations. (From Otto Nathan and Heinz Norden [eds.], *Einstein on Peace;* Simon and Schuster, 1960.)

It would be said often of Einstein that he was naïve about human affairs; for example, with the proclamation of the German Republic and the armistice in 1918, he was convinced that militarism had been thoroughly abolished in Germany.

International acclaim. International fame came to Einstein in November 1919, when the Royal Society of London announced that its scientific expedition to Príncipe Island, in the Gulf of Guinea, had photographed the solar eclipse on May 29 of that year and completed calculations that verified the predictions made in Einstein's general theory of relativity. Few could understand relativity, but the basic postulates were so revolutionary and the scientific community was so obviously bedazzled that the physicist was acclaimed the greatest genius on Earth. Einstein himself was amazed at the reaction and apparently displeased, for he resented the consequent interruptions of his work. After his divorce he had, in the summer of 1919, married Elsa, the widowed daughter of his late father's cousin. He lived quietly with Elsa and her two daughters in Berlin, but, inevitably, his views as a foremost savant were sought on a variety of issues.

Proof of the general theory of relativity

Despite the now deteriorating political situation in Germany, Einstein attacked nationalism and promoted pacifist ideals. With the rising tide of anti-Semitism in Berlin, Einstein was castigated for his "Bolshevism in physics," and the fury against him in right-wing circles grew when he began publicly to support the Zionist movement. Judaism had played little part in his life, but he insisted that, as a snail can shed his shell and still be a snail, so a Jew can shed his faith and still be a Jew.

Although Einstein was regarded warily in Berlin, such was the demand for him in other European cities that he travelled widely to lecture on relativity, usually arriving at each place by third-class rail carriage, with a violin tucked under his arm. So successful were his lectures that one enthusiastic impresario guaranteed him a three-week booking at the London Palladium. He ignored the offer, but, at the request of the Zionist leader Chaim Weizmann, toured the United States in the spring of 1921 to raise money for the Palestine Foundation Fund. Frequently treated like a circus freak and feted from morning to night, Einstein nevertheless was gratified by the standards of scientific research and the "idealistic attitudes" that he found prevailing in the United States.

During the next three years Einstein was constantly on the move, journeying not only to European capitals but also to the Orient, to the Middle East, and to South America. According to his diary notes, he found nobility among the Hindus of Ceylon, a pureness of soul among the Japanese, and a magnificent intellectual and moral calibre among the Jewish settlers in Palestine. His wife later wrote that, on steaming into one new harbour, Einstein had said to her, "Let us take it all in before we wake up."

In Shanghai a cable reached him announcing that he had been awarded the 1921 Nobel Prize for Physics "for your photoelectric law and your work in the field of theoretical physics." Relativity, still the centre of controversy, was not mentioned.

The Nobel Prize

Though the 1920s were tumultuous times of wide acclaim, and some notoriety, Einstein did not waver from his new search—to find the mathematical relationship between electromagnetism and gravitation. This would be a first step, he felt, in discovering the common laws governing the behaviour of everything in the universe, from the electron to the planets. He sought to relate the universal properties of matter and energy in a single equation or formula, in what came to be called a unified field theory. This turned out to be a fruitless quest that occupied the rest of his life. Einstein's peers generally agreed quite early that his search was destined to fail because the rapidly developing quantum theory uncovered an uncertainty principle in all measurements of the motion of particles: the movement of a single particle simply could not be predicted because of a fundamental uncertainty in measuring simultaneously both its speed and its position, which means, in effect, that the future of any physical system at the subatomic level cannot be predicted. While fully recognizing the brilliance of quantum mechanics, Einstein rejected the idea that these theories were absolute and persevered with his theory of general relativity as the more satisfactory foundation to future discovery. He was widely quoted on his belief in an exactly engineered universe: "God is subtle but he is not malicious." On this point, he parted company with most theoretical physicists. The distinguished German quantum theorist Max Born, a close friend of Einstein, said at the time: "Many of us regard this as a tragedy, both for him, as he gropes his way in loneliness, and for us, who miss our leader and standard-bearer." This appraisal, and others pronouncing his work in later life as largely wasted effort, will have to await the judgment of later generations.

The year of Einstein's 50th birthday, 1929, marked the beginning of the ebb flow of his life's work in a number of aspects. Early in the year the Prussian Academy published the first version of his unified-field theory, but, despite the sensation it caused, its very preliminary nature soon became apparent. The reception of the theory left him undaunted, but Einstein was dismayed by the preludes to certain disaster in the field of human affairs: Arabs launched savage attacks on Jewish colonists in Palestine; the Nazis gained strength in Germany; the League of Nations proved so impotent that Einstein resigned abruptly from its Committee on Intellectual Cooperation as a protest to its timidity; and the stock market crash in New York City heralded worldwide economic crisis.

Crushing Einstein's natural gaiety more than any of these events was the mental breakdown of his younger son, Edward. Edward had worshipped his father from a distance but now blamed him for deserting him and for ruining his life. Einstein's sorrow was eased only slightly by the amicable relationship he enjoyed with his older son, Hans Albert.

As visiting professor at Oxford University in 1931, Ein-

stein spent as much time espousing pacifism as he did discussing science. He went so far as to authorize the establishment of the Einstein War Resisters' International Fund in order to bring massive public pressure to bear on the World Disarmament Conference, scheduled to meet in Geneva in February 1932. When these talks foundered, Einstein felt that his years of supporting world peace and human understanding had accomplished nothing. Bitterly disappointed, he visited Geneva to focus world attention on the "farce" of the disarmament conference. In a rare moment of fury, Einstein stated to a journalist,

Einstein War Resisters' International Fund

> They [the politicians and statesmen] have cheated us. They have fooled us. Hundreds of millions of people in Europe and in America, billions of men and women yet to be born, have been and are being cheated, traded and tricked out of their lives and health and well-being.

Shortly after this, in a famous exchange of letters with the Austrian psychiatrist Sigmund Freud, Einstein suggested that people must have an innate lust for hatred and destruction. Freud agreed, adding that war was biologically sound because of the love–hate instincts of man and that pacifism was an idiosyncrasy directly related to Einstein's high degree of cultural development. This exchange was only one of Einstein's many philosophic dialogues with renowned men of his age. With Rabindranath Tagore, Hindu poet and mystic, he discussed the nature of truth. While Tagore held that truth was realized through man, Einstein maintained that scientific truth must be conceived as a valid truth that is independent of humanity. "I cannot prove that I am right in this, but that is my religion," said Einstein. Firmly denying atheism, Einstein expressed a belief in "Spinoza's God who reveals himself in the harmony of what exists." The physicist's breadth of spirit and depth of enthusiasm were always most evident among truly intellectual men. He loved being with the physicists Paul Ehrenfest and Hendrick A. Lorentz at The Netherlands' Leiden University, and several times he visited the California Institute of Technology in Pasadena to attend seminars at the Mt. Wilson Observatory (now part of Hale Observatories), which had become world renowned for astrophysical research. At Mt. Wilson he heard the Belgian scientist Abbé Georges Lemaître detail his theory that the universe had been created by the explosion of a "primeval atom" and was still expanding. Gleefully, Einstein jumped to his feet, applauding. "This is the most beautiful and satisfactory explanation of creation to which I have ever listened," he said.

In 1933, soon after Adolf Hitler became chancellor of Germany, Einstein renounced his German citizenship and left the country. He later accepted a full-time position as a foundation member of the school of mathematics at the new Institute for Advanced Study in Princeton, New Jersey. In reprisal, Nazi storm troopers ransacked his beloved summer house at Caputh, near Berlin, and confiscated his sailboat. Einstein was so convinced that Nazi Germany was preparing for war that, to the horror of Romain Rolland and his other pacifist friends, he violated his pacifist ideals and urged free Europe to arm and recruit for defense.

Flight from Europe

Although his warnings about war were largely ignored, there were fears for Einstein's life. He was taken by private yacht from Belgium to England. By the time he arrived in Princeton in October 1933, he had noticeably aged. A friend wrote,

> It was as if something had deadened in him. He sat in a chair at our place, twisting his white hair in his fingers and talking dreamily about everything under the sun. He was not laughing any more.

Later years in the United States. In Princeton Einstein set a pattern that was to vary little for more than 20 years. He lived with his wife in a simple, two-story frame house and most mornings walked a mile or so to the Institute, where he worked on his unified-field theory and talked with colleagues. For relaxation he played his violin and sailed on a local lake. Only rarely did he travel, even to New York. In a letter to Queen Elisabeth of Belgium, he described his new refuge as a "wonderful little spot, . . . a quaint and ceremonious village of puny demigods on stilts." Eventually he acquired American citizenship, but he always continued to think of himself as a European. Pursuing his own line of theoretical research outside the mainstream of physics, he took on an air of fixed serenity. "Among my European friends, I am now called *Der grosse Schweiger* ("The Great Stone Face"), a title I well deserve," he said. Even his wife's death late in 1936 did not disturb his outward calm. "It seemed that the difference between life and death for Einstein consisted only in the difference between being able and not being able to do physics," wrote Leopold Infeld, the Polish physicist who arrived in Princeton at this time.

Development of the atomic bomb

Niels Bohr, the great Danish atomic physicist, brought news to Einstein in 1939 that the German refugee physicist Lise Meitner had split the uranium atom, with a slight loss of total mass that had been converted into energy. Meitner's experiments, performed in Copenhagen, had been inspired by similar, though less precise, experiments done months earlier in Berlin by two German chemists, Otto Hahn and Fritz Strassmann. Bohr speculated that, if a controlled chain-reaction splitting of uranium atoms could be accomplished, a mammoth explosion would result. Einstein was skeptical, but laboratory experiments in the United States showed the feasibility of the idea. With a European war regarded as imminent and fears that Nazi scientists might build such a "bomb" first, Einstein was persuaded by colleagues to write a letter to President Franklin D. Roosevelt urging "watchfulness and, if necessary, quick action" on the part of the United States in atomic-bomb research. This recommendation marked the beginning of the Manhattan Project.

Although he took no part in the work at Los Alamos, New Mexico, and did not learn that a nuclear-fission bomb had been made until Hiroshima was razed in 1945, Einstein's name was emphatically associated with the advent of the atomic age. He readily joined those scientists seeking ways to prevent any future use of the bomb, his particular and urgent plea being the establishment of a world government under a constitution drafted by the United States, Britain, and Russia. With the spur of the atomic fear that haunted the world, he said "we must not be merely willing, but actively eager to submit ourselves to the binding authority necessary for world security." Once more, Einstein's name surged through the newspapers. Letters and statements tumbled out of his Princeton study, and in the public eye Einstein the physicist dissolved into Einstein the world citizen, a kind "grand old man" devoting his last years to bringing harmony to the world.

The rejection of his ideals by statesmen and politicians did not break him, because his prime obsession still remained with physics. "I cannot tear myself away from my work," he wrote at the time. "It has me inexorably in its clutches." In proof of this came his new version of the unified field in 1950, a most meticulous mathematical essay that was immediately but politely criticized by most physicists as untenable.

New version of the unified field

Compared with his renown of a generation earlier, Einstein was virtually neglected and said himself that he felt almost like a stranger in the world. His health deteriorated to the extent that he could no longer play the violin or sail his boat. Many years earlier, chronic abdominal pains had forced him to give up smoking his pipe and to watch his diet carefully.

On April 18, 1955, Einstein died in his sleep at Princeton Hospital. On his desk lay his last incomplete statement, written to honour Israeli Independence Day. It read in part: "What I seek to accomplish is simply to serve with my feeble capacity truth and justice at the risk of pleasing no one." His contribution to man's understanding of the universe was matchless, and he is established for all time as a giant of science. Broadly speaking, his crusades in human affairs seem to have had no lasting impact. Einstein perhaps anticipated such an assessment of his life when he said, "Politics are for the moment. An equation is for eternity." (P.Mi.)

MAJOR WORKS

SCIENTIFIC PAPERS: "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," in *Annalen der Physik* (1905); "Über die von der molekularkinetis-

chen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," in *Annalen der Physik* (1905); "Zur Elektrodynamik bewegter Körper," in *Annalen der Physik* (1905), the initial paper on special relativity; "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?" in *Annalen der Physik* (1905); "Zur Theorie der Brownschen Bewegung," in *Annalen der Physik* (1906), translated separately as *Investigations on the Theory of the Brownian Movement* (1926); "Zur Theorie der Lichterzeugung und Lichtabsorption," in *Annalen der Physik* (1906); "Plancksche Theorie der Strahlung und die Theorie der spezifischen Wärme," in *Annalen der Physik* (1907); "Entwurf einer Verallegemeinerten Relativitätstheorie und einer Theorie der Gravitation," in *Zeitschrift für Mathematik und Physik* (1913); "Grundlagen der allgemeinen Relativitätstheorie," in *Annalen der Physik* (1916), on the general theory of relativity; "Strahlungs-emission und -absorption nach der Quantentheorie," in *Verhandlungen der Deutschen physikalischen Gesellschaft* (1916); "Quantentheorie der Strahlung," in *Physikalische Zeitschrift* (1917); "Quantentheorie des einatomigen idealen Gases," in *Sitzungsberichte der Preussischen Akademie der Wissenschaften* (1924 and 1925). Some of Einstein's important papers were collected in the joint work (with H.A. Lorentz and H. Minkowski), *H.A. Lorentz: Das Relativitätsprinzip, eine Sammlung von Abhandlungen* (1913; trans. as *H.A. Lorentz: The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity,* 1923). See also *The Meaning of Relativity,* which includes the generalized theory of gravitation (1953), the first edition of Einstein's unified-field theory.

OTHER WORKS: *About Zionism: Speeches and Letters,* Eng. trans. by Sir Leon Simon (1931); *Builders of the Universe* (1932); with Sigmund Freud, *Warum Krieg? (Why War?,* Eng. trans. by Stuart Gilbert, 1933); with Leopold Infeld, *The Evolution of Physics* (1938); *The World As I See It* (Eng. trans. by Alan Harris, 1949); *Out of My Later Years* (1950).

BIBLIOGRAPHY. PAUL A. SCHILPP (ed.), *Albert Einstein: Philosopher-Scientist,* 3rd ed., 2 vol. (1969), a discussion by eminent scholars of Einstein's impact on science and philosophy; PHILIPP FRANK, *Einstein: His Life and Times* (1947, reissued 1972), a scientific biography on Einstein's early life and achievement; ANTONINA VALLENTIN, *Einstein: A Biography* (U.S. title, *The Drama of Albert Einstein,* 1954; trans. from the French, 1954), a personal story of Einstein's European years; PETER MICHELMORE, *Einstein: Profile of the Man* (1962), a popular biography, richly anecdotal, of Einstein as man and scientist; LINCOLN BARNETT, *The Universe and Dr. Einstein,* 2nd rev. ed. (1957, reissued 1974), a lucid exposition of Einstein's contribution to science; RONALD W. CLARK, *Einstein: The Life and Times* (1971, reissued 1979), a distinguished, definitive work (well illustrated); BANESH HOFFMAN, with the collaboration of HELEN DUKAS, *Albert Einstein: Creator and Rebel* (1972), a significant biography, laced with a thorough but exciting interpretation of Einstein's scientific work; JEREMY BERNSTEIN, *Einstein* (1973), a biography emphasizing the scientific theories; CORNELIUS LANCZOS, *The Einstein Decade: 1905–1915* (1974), a biography that includes detailed synopses of each Einstein paper written during the years covered; HELEN DUKAS and BANESH HOFFMAN (eds.), *Albert Einstein, the Human Side: New Glimpses from His Archives* (1979), a sampling of his letters that provides a good introduction to his personality and thought; A.P. FRENCH (ed.), *Einstein: A Centenary Volume* (1979), a collection of essays, reminiscences, illustrations, and quotations—for the general audience. ABRAHAM PAIS, *'Subtle is the Lord. . .': The Science and the Life of Albert Einstein* (1982), a scientific biography.

Electricity and Magnetism

Electricity and magnetism are two aspects of electromagnetism, the science of charge and of the forces and fields associated with charge. Electricity and magnetism were long thought to be separate forces. It was not until the 19th century that they were finally treated as interrelated phenomena. In 1905 Albert Einstein's special theory of relativity established beyond a doubt that both are aspects of one common phenomenon. At a practical level, however, electric and magnetic forces behave quite differently and are described by different equations. Electric forces are produced by electric charges either at rest or in motion. Magnetic forces, on the other hand, are produced only by moving charges and act solely on charges in motion.

Electric phenomena occur even in neutral matter because the forces act on the individual charged constituents. The electric force, in particular, is responsible for most of the physical and chemical properties of atoms and molecules. It is enormously strong compared with gravity. For example, the absence of only one electron out of every billion molecules in two 70-kilogram (154-pound) persons standing two metres (two yards) apart would repel them with a 30,000-ton force. On a more familiar scale, electric phenomena are responsible for the lightning and thunder accompanying certain storms.

Electric and magnetic forces can be detected in regions called electric and magnetic fields. These fields are fundamental in nature and can exist in space far from the charge or current that generated them. Remarkably, electric fields can produce magnetic fields and vice versa, independent of any external charge. A changing magnetic field produces an electric field, as the English physicist Michael Faraday discovered in work that forms the basis of electric power generation. Conversely, a changing electric field produces a magnetic field, as the Scottish physicist James Clerk Maxwell deduced. The mathematical equations formulated by Maxwell incorporated light and wave phenomena into electromagnetism. He showed that electric and magnetic fields travel together through space as waves of electromagnetic radiation, with the changing fields mutually sustaining each other. Examples of electromagnetic waves traveling through space independent of matter are radio and television waves, microwaves, infrared rays, visible light, ultraviolet light, X rays, and gamma rays. All of these waves travel at the same speed—namely, the velocity of light (roughly 300,000 kilometres, or 186,000 miles, per second). They differ from each other only in the frequency at which their electric and magnetic fields oscillate.

Maxwell's equations still provide a complete and elegant description of electromagnetism down to, but not including, the subatomic scale. The interpretation of his work, however, was broadened in the 20th century. Einstein's special relativity theory merged electric and magnetic fields into one common field and limited the velocity of all matter to the velocity of electromagnetic radiation. During the late 1960s, physicists discovered that other forces in nature have fields with a mathematical structure similar to that of the electromagnetic field. These other forces are the nuclear force, responsible for the energy released in nuclear fusion, and the weak force, observed in the radioactive decay of unstable atomic nuclei. In particular, the weak and electromagnetic forces have been combined into a common force called the electroweak force. The goal of many physicists to unite all of the fundamental forces, including gravity, into one grand unified theory has not been attained to date.

An important aspect of electromagnetism is the science of electricity, which is concerned with the behaviour of aggregates of charge, including the distribution of charge within matter and the motion of charge from place to place. Different types of materials are classified as either conductors or insulators on the basis of whether charges can move freely through their constituent matter. Electric current is the measure of the flow of charges; the laws governing currents in matter are important in technology, particularly in the production, distribution, and control of energy.

The concept of voltage, like those of charge and current, is fundamental to the science of electricity. Voltage is a measure of the propensity of charge to flow from one place to another; positive charges generally tend to move from a region of high voltage to a region of lower voltage. A common problem in electricity is determining the relationship between voltage and current or charge in a given physical situation.

This article seeks to provide a qualitative understanding of electromagnetism as well as a quantitative appreciation for the magnitudes associated with electromagnetic phenomena.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 127 and 128, and the *Index.*

The article is divided into the following sections:

General considerations

Everyday modern life is pervaded by electromagnetic phenomena. When a light bulb is switched on, a current flows through a thin filament in the bulb; the current heats the filament to such a high temperature that it glows, illuminating its surroundings. Electric clocks and connections link simple devices of this kind into complex systems such as traffic lights that are timed and synchronized with the speed of vehicular flow. Radio and television sets receive information carried by electromagnetic waves traveling through space at the speed of light. To start an automobile, currents in an electric starter motor generate magnetic fields that rotate the motor shaft and drive engine pistons to compress an explosive mixture of gasoline and air; the spark initiating the combustion is an electric discharge, which makes up a momentary current flow.

Many of these devices and phenomena are complex, but they derive from the same fundamental laws of electromagnetism. One of the most important of these is Coulomb's law, which describes the electric force between charged objects. Formulated by the 18th-century French physicist Charles-Augustin de Coulomb, it is analogous to Newton's law for the gravitational force. Both gravitational and electric forces decrease with the square of the distance between the objects, and both forces act along a line between them. In Coulomb's law, however, the magnitude and sign of the electric force are determined by the charge, rather than the mass, of an object. Thus, charge determines how electromagnetism influences the motion of charged objects. (Charge is a basic property of matter. Every constituent of matter has an electric charge with a value that can be positive, negative, or zero. For example, electrons are negatively charged, and atomic nuclei are positively charged. Most bulk matter has an equal amount of positive and negative charge and thus has zero net charge.)

Coulomb's law

According to Coulomb, the electric force for charges at rest has the following properties:

(1) Like charges repel each other; unlike charges attract. Thus, two negative charges repel one another, while a positive charge attracts a negative charge.

(2) The attraction or repulsion acts along the line between the two charges.

(3) The size of the force varies inversely as the square of the distance between the two charges. Therefore, if the distance between the two charges is doubled, the attraction or repulsion becomes weaker, decreasing to one-fourth of the original value. If the charges come 10 times closer, the size of the force increases by a factor of 100.

(4) The size of the force is proportional to the value of each charge. The unit used to measure charge is the coulomb (C). If there were two positive charges, one of 0.1 coulomb and the second of 0.2 coulomb, they would repel each other with a force that depends on the product 0.2×0.1. If each of the charges were reduced by one-half, the repulsion would be reduced to one-quarter of its former value.

Static cling is a practical example of the Coulomb force. In static cling, garments made of synthetic material collect a charge, especially in dry winter air. A plastic or rubber comb passed quickly through hair also becomes charged and will pick up bits of paper. The synthetic fabric and the comb are insulators; charge on these objects cannot move easily from one part of the object to another. Similarly, an office copy machine uses electric force to attract particles of ink to paper.

Like Coulomb's law, the principle of charge conservation is a fundamental law of nature. According to this principle, the charge of an isolated system cannot change. If an additional positively charged particle appears within a system, a particle with a negative charge of the same magnitude will be created at the same time; thus, the principle of conservation of charge is maintained. In nature, a pair of oppositely charged particles is created when high-energy radiation interacts with matter; an electron and a positron are created in a process known as pair production.

Principle of charge conservation

The smallest subdivision of the amount of charge that a particle can have is the charge of one proton, $+1.602 \times 10^{-19}$ coulomb. The electron has a charge of the same magnitude but opposite sign—*i.e.*, -1.602×10^{-19} coulomb. An ordinary flashlight battery delivers a current that provides a total charge flow of approximately 5,000 coulomb, which corresponds to more than 10^{22} electrons, before it is exhausted.

Electric current is a measure of the flow of charge, as, for example, charge flowing through a wire. The size of the current is measured in amperes and symbolized by *i*. An ampere of current represents the passage of one coulomb of charge per second, or 6.2 billion billion electrons (6.2×10^{18} electrons) per second. A current is positive when it is in the direction of the flow of positive charges; its direction is opposite to the flow of negative charges.

The force and conservation laws are only two aspects of electromagnetism, however. Electric and magnetic forces are caused by electromagnetic fields. The term field denotes a property of space, so that the field quantity has a numerical value at each point of space. These values may also vary with time. The value of the electric or magnetic field is a vector—*i.e.*, a quantity having both magnitude and direction. The value of the electric field at a point in space, for example, equals the force that would be exerted on a unit charge at that position in space.

Every charged object sets up an electric field in the surrounding space. A second charge "feels" the presence of this field. The second charge is either attracted toward the initial charge or repelled from it, depending on the signs of the charges. Of course, since the second charge also has an electric field, the first charge feels its presence and is either attracted or repelled by the second charge, too.

Electric fields and forces

The electric field from a charge is directed away from the charge when the charge is positive and toward the charge when it is negative. The electric field from a charge at rest is shown in Figure 1 for various locations in space. The arrows point in the direction of the electric field, and the length of the arrows indicates the strength of the field at the midpoint of the arrows.

If a positive charge were placed in the electric field, it would feel a force in the direction of the field. A negative charge would feel a force in the direction opposite the direction of the field.

In calculations, it is often more convenient to deal directly with the electric field than with the charges; fre-

By courtesy of the Department of Physics and Astronomy, Michigan State University

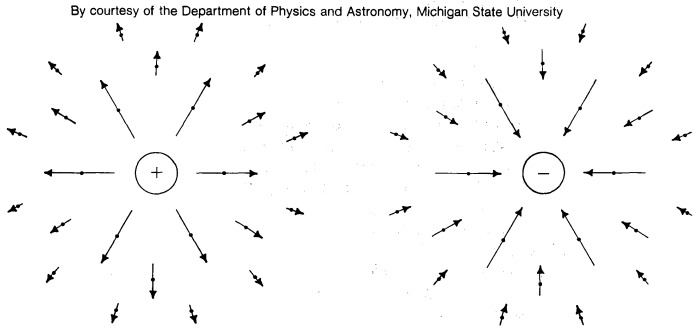


Figure 1: *Electric fields.*
(Left) Field of a positive electric charge; (right) field of a negative electric charge.

quently, more is known about the field than about the distribution of charges in space. For example, the distribution of charges in conductors is generally unknown because the charges move freely within the conductor. In static situations, however, the electric field in a conductor in equilibrium has a definite value, zero, because any force on the charges inside the conductor redistributes them until the field vanishes. The unit of electric field is newtons per coulomb, or volts per metre.

Electric potential

The electric potential is another useful field. It provides an alternative to the electric field in electrostatics problems. The potential is easier to use, however, because it is a single number, a scalar, instead of a vector. The difference in potential between two places measures the degree to which charges are influenced to move from one place to another. If the potential is the same at two places (*i.e.*, if the places have the same voltage), charges will not be influenced to move from one place to the other. The potential on an object or at some point in space is measured in volts; it equals the electrostatic energy that a unit charge would have at that position. In a typical 12-volt car battery, the battery terminal that is marked with a + sign is at a potential 12 volts greater than the potential of the terminal marked with the − sign. When a wire, such as the filament of a car headlight, is connected between the + and the − terminals of the battery, charges move through the filament as an electric current and heat the filament; the hot filament radiates light.

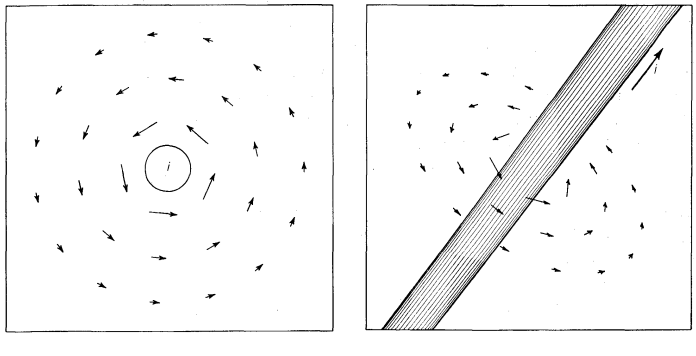
Magnetic fields and forces

The magnetic force influences only those charges that are already in motion. It is transmitted by the magnetic field. Both magnetic fields and magnetic forces are more complicated than electric fields and electric forces. The magnetic field does not point along the direction of the source of the field; instead, it points in a perpendicular direction. In addition, the magnetic force acts in a direction that is perpendicular to the direction of the field. In comparison, both the electric force and the electric field point directly toward or away from the charge.

The present discussion will deal with simple situations in which the magnetic field is produced by a current of charge in a wire. Certain materials, such as copper, silver, and aluminum, are conductors that allow charge to flow freely from place to place. If an external influence establishes a current in a conductor, the current generates a magnetic field. For a long straight wire, the magnetic field has a direction that encircles the wire on a plane perpendicular to the wire. The strength of the magnetic field decreases with distance from the wire. The arrows in Figure 2 represent the size and direction of the magnetic field for a current moving in the direction indicated. Figure 2A shows an end view with the current coming toward the reader, while Figure 2B provides a three-dimensional view of the magnetic field at one position along the wire.

In subsequent figures, continuous lines will be used to represent the direction of electric and magnetic fields. These lines emphasize the important fact that electric fields begin on positive charges and end on negative charges, while magnetic fields do not have beginnings or ends and close on themselves. The magnetic field shown in Figure 2

By courtesy of the Department of Physics and Astronomy, Michigan State University



A B

Figure 2: *Magnetic field of a long wire.*
(A) An end view, with the current flowing toward the reader.
(B) A three-dimensional view.

is unusually simple. Highly complex and useful magnetic fields can be generated by the proper choice of conductors to carry electric currents. Under development are thermonuclear fusion reactors for obtaining energy from the fusion of light nuclei in the form of very hot plasmas of hydrogen isotopes. The plasmas have to be confined by magnetic fields (dubbed "magnetic bottles") as no material container can withstand such high temperatures. Charged particles are also confined by magnetic fields in nature. Large numbers of charged particles, mostly protons and electrons, are trapped in huge bands around the Earth by its magnetic field. These bands are known as the Van Allen radiation belts. Disturbance of the Earth's confining magnetic field produces spectacular displays, the so-called northern lights, in which trapped charged particles are freed and crash through the atmosphere to Earth.

Interaction of a magnetic field with charge

How does the magnetic field interact with a charged object? If the charge is at rest, there is no interaction. If the charge moves, however, it is subjected to a force, the size of which increases in direct proportion with the velocity of the charge. The force has a direction that is perpendicular both to the direction of motion of the charge and to the direction of the magnetic field. There are two possible precisely opposite directions for such a force for a given direction of motion. This apparent ambiguity is resolved by the fact that one of the two directions applies to the force on a moving positive charge while the other direction applies to the force on a moving negative charge. Figure 3 illustrates the directions of the magnetic force on positive charges and on negative charges as they move in a magnetic field that is perpendicular to the motion.

By courtesy of the Department of Physics and Astronomy, Michigan State University

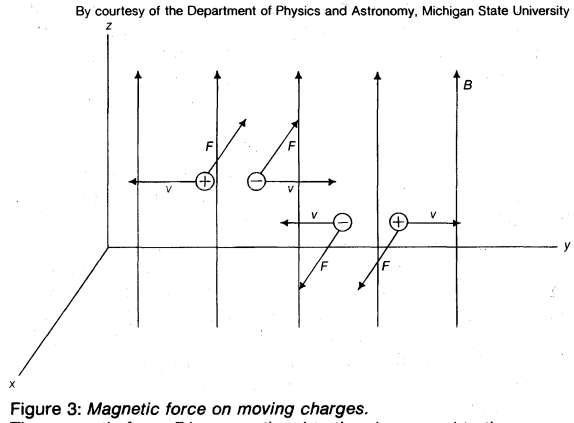


Figure 3: *Magnetic force on moving charges.*
The magnetic force ***F*** is proportional to the charge and to the magnitude of velocity ***v*** times the magnetic field ***B***.

Depending on the initial orientation of the particle velocity to the magnetic field, charges having a constant speed in a uniform magnetic field will follow a circular or helical path.

Electric currents in wires are not the only source of magnetic fields. Naturally occurring minerals exhibit magnetic properties and have magnetic fields. These magnetic fields result from the motion of electrons in the atoms of the material. They also result from a property of electrons called the magnetic dipole moment, which is related to the intrinsic spin of individual electrons (see the article ATOMS: *Electrons*). In most materials, little or no field is observed outside the matter because of the random orientation of the various constituent atoms. In some materials such as iron, however, atoms within certain distances tend to become aligned in one particular direction.

Magnets have numerous applications, ranging from use as toys and paper holders on home refrigerators to essential components in electric generators and machines that can accelerate particles to speeds approaching that of light. The practical application of magnetism in technology is greatly enhanced by using iron and other ferromagnetic materials with electric currents in devices like motors. These materials amplify the magnetic field produced by the currents and thereby create more powerful fields (see below *Ferromagnetism*).

While electric and magnetic effects are well separated in many phenomena and applications, they are coupled

Time-varying magnetic field

closely together when there are rapid time fluctuations. Faraday's law of induction describes how a time-varying magnetic field produces an electric field (see below *Faraday's law of induction*). Important practical applications include the electric generator and transformer. In a generator, the physical motion of a magnetic field produces electricity for power. In a transformer, electric power is converted from one voltage level to another by the magnetic field of one circuit inducing an electric current in another circuit.

The existence of electromagnetic waves depends on the interaction between electric and magnetic fields. Maxwell postulated that a time-varying electric field produces a magnetic field. His theory predicted the existence of electromagnetic waves in which each time-varying field produces the other field. For example, radio waves are generated by electronic circuits known as oscillators that cause rapidly oscillating currents to flow in antennas; the rapidly varying magnetic field has an associated varying electric field. The result is the emission of radio waves into space (see ELECTROMAGNETIC RADIATION).

Many electromagnetic devices can be described by circuits consisting of conductors and other elements. These circuits may operate with a steady flow of current, as in a flashlight, or with time-varying currents. Important elements in circuits include sources of power called electromotive forces; resistors, which control the flow of current for a given voltage; capacitors, which store charge and energy temporarily; and inductors, which also store electrical energy for a limited time. Circuits with these elements can be described entirely with algebra. (For more complicated circuit elements such as transistors, see ELECTRONICS: *Semiconductor devices* and *Integrated circuits*).

Fluxes and line integrals of vector fields

Two mathematical quantities associated with vector fields, like the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$, are useful for describing electromagnetic phenomena. They are the flux of such a field through a surface and the line integral of the field along a path. The flux of a field through a surface measures how much of the field penetrates through the surface; for every small section of the surface, the flux is proportional to the area of that section and depends also on the relative orientation of the section and the field. The line integral of a field along a path measures the degree to which the field is aligned with the path; for every small section of path, it is proportional to the length of that section and is also dependent on the alignment of the field with that section of path. When the field is perpendicular to the path, there is no contribution to the line integral. The fluxes of $\boldsymbol{E}$ and $\boldsymbol{B}$ through a surface and the line integrals of these fields along a path play an important role in electromagnetic theory. As examples, the flux of the electric field $\boldsymbol{E}$ through a closed surface measures the amount of charge contained within the surface; the flux of the magnetic field $\boldsymbol{B}$ through a closed surface is always zero because there are no magnetic monopoles (magnetic charges consisting of a single pole) to act as sources of the magnetic field in the way that charge is a source of the electric field.

Electricity

ELECTROSTATICS

Use of mathematical methods

Electrostatics is the study of electromagnetic phenomena that occur when there are no moving charges—*i.e.*, after a static equilibrium has been established. Charges reach their equilibrium positions rapidly because the electric force is extremely strong. The mathematical methods of electrostatics make it possible to calculate the distributions of the electric field and of the electric potential from a known configuration of charges, conductors, and insulators. Conversely, given a set of conductors with known potentials, it is possible to calculate electric fields in regions between the conductors and to determine the charge distribution on the surface of the conductors. The electric energy of a set of charges at rest can be viewed from the standpoint of the work required to assemble the charges; alternatively, the energy also can be considered to reside in the electric field produced by this assembly of charges. Finally, energy can be stored in a capacitor; the energy required to charge such a device is stored in it as electrostatic energy of the electric field.

Static electricity. This is a familiar electric phenomenon in which friction transfers charged particles from one body to another. If two objects are rubbed together, especially if the objects are insulators and the surrounding air is dry, the objects acquire equal and opposite charges and an attractive force develops between them. The object that loses electrons becomes positively charged, and the other becomes negatively charged. The force is simply the attraction between charges of opposite sign. The properties of this force were described above; they are incorporated in the mathematical relationship known as Coulomb's law. The electric force on a charge Q_1 under these conditions, due to a charge Q_2 at a distance r, is given by Coulomb's law,

$$\boldsymbol{F} = k\frac{Q_1 Q_2}{r^2}\hat{\boldsymbol{r}} \qquad (1)$$

The bold characters in the equation indicate the vector nature of the force, and the unit vector $\hat{\boldsymbol{r}}$ is a vector that has a size of one and that points from charge Q_2 to charge Q_1. The proportionality constant k equals $10^{-7}c^2$, where c is the speed of light in a vacuum; k has the numerical value of 8.99×10^9 newtons-square metre per coulomb squared (Nm^2/C^2). Figure 4 shows the force on Q_1 due to Q_2. A numerical example will help to illustrate this force. Both Q_1 and Q_2 are chosen arbitrarily to be positive charges, each with a magnitude of 10^{-6} coulomb. The charge Q_1 is located at coordinates x, y, z with values of 0.03, 0, 0, respectively, while Q_2 has coordinates 0, 0.04, 0. All coordinates are given in metres. Thus, the distance between Q_1 and Q_2 is 0.05 metre.

By courtesy of the Department of Physics and Astronomy, Michigan State University

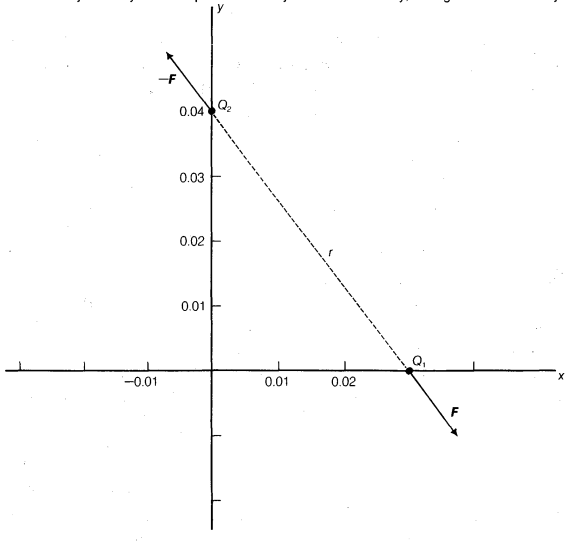


Figure 4: Electric force between two charges (see text).

The magnitude of the force $\boldsymbol{F}$ on charge Q_1 as calculated using equation (1) is 3.6 newtons; its direction is shown in Figure 4. The force on Q_2 due to Q_1 is $-\boldsymbol{F}$, which also has a magnitude of 3.6 newtons; its direction, however, is opposite to that of $\boldsymbol{F}$. The force $\boldsymbol{F}$ can be expressed in terms of its components along the x and y axes, since the force vector lies in the xy plane. This is done with elementary trigonometry from the geometry of Figure 4, and the results are shown in Figure 5. Thus,

$$\boldsymbol{F} = 2.16\hat{\boldsymbol{x}} - 2.88\hat{\boldsymbol{y}} \qquad (2)$$

in newtons. Coulomb's law describes mathematically the properties of the electric force between charges at rest. If the charges have opposite signs, the force would be attractive; the attraction would be indicated in equation (1) by the negative coefficient of the unit vector $\hat{\boldsymbol{r}}$. Thus, the electric force on Q_1 would have a direction opposite to the unit vector $\hat{\boldsymbol{r}}$ and would point from Q_1 to Q_2. In Cartesian coordinates, this would result in a change of the signs of both the x and y components of the force in equation (2).

How can this electric force on Q_1 be understood? Fun-

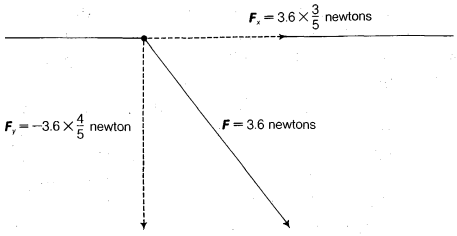


Figure 5: The x and y components of the force F in Figure 4 (see text).

By courtesy of the Department of Physics and Astronomy, Michigan State University

damentally, the force is due to the presence of an electric field at the position of Q_1. The field is caused by the second charge Q_2 and has a magnitude proportional to the size of Q_2. In interacting with this field, the first charge some distance away is either attracted to or repelled from the second charge, depending on the sign of the first charge.

Calculating the value of an electric field

In the example, the charge Q_1 is in the electric field produced by the charge Q_2. This field has the value

$$E = k\frac{Q_2}{r^2}\hat{r} \qquad (3)$$

in newtons per coulomb (N/C). (Electric field can also be expressed in volts per metre [V/m], which is the equivalent of newtons per coulomb.) The electric force on Q_1 is given by

$$F = Q_1 E \qquad (4)$$

in newtons. This equation can be used to define the electric field of a point charge. The electric field E produced by charge Q_2 is a vector. The magnitude of the field varies inversely as the square of the distance from Q_2; its direction is away from Q_2 when Q_2 is a positive charge and toward Q_2 when Q_2 is a negative charge. Using equations (2) and (4), the field produced by Q_2 at the position of Q_1 is

$$E = 2.16 \times 10^6\hat{x} - 2.88 \times 10^6\hat{y}$$

in newtons per coulomb.

When there are several charges present, the force on a given charge Q_1 may be simply calculated as the sum of the individual forces due to the other charges Q_2, Q_3, . . . , etc., until all the charges are included. This sum requires that special attention be given to the direction of the individual forces since forces are vectors. The force on Q_1 can be obtained with the same amount of effort by first calculating the electric field at the position of Q_1 due to Q_2, Q_3, . . . , etc. To illustrate this, a third charge is added to the example above. There are now three charges, $Q_1 = +10^{-6}$ C, $Q_2 = +10^{-6}$ C, and $Q_3 = -10^{-6}$ C. The locations of the charges, using Cartesian coordinates [x, y, z] are, respectively, [0.03, 0, 0], [0, 0.04, 0], and [−0.02, 0, 0] metre, as shown in Figure 6. The goal is to find the force on Q_1. From the sign of the charges, it can be seen that Q_1 is repelled by Q_2 and attracted by Q_3. It is also clear that these two forces act along different directions. The electric field at the position of Q_1 due to charge Q_2 is, just as in the example above,

$$E_{1,2} = 2.16 \times 10^6\hat{x} - 2.88 \times 10^6\hat{y}$$

in newtons per coulomb. The electric field at the location of Q_1 due to charge Q_3 is

$$E_{1,3} = -3.6 \times 10^6\hat{x}$$

in newtons per coulomb. Thus, the total electric field at position 1 (*i.e.*, at [0.03, 0, 0]) is the sum of these two fields $E_{1,2} + E_{1,3}$ and is given by

$$E_1\ (total) = -1.44 \times 10^6\hat{x} - 2.88 \times 10^6\hat{y}.$$

The fields $E_{1,2}$ and $E_{1,3}$, as well as their sum, the total electric field at the location of Q_1, E_1 (*total*), are shown in Figure 6. The total force on Q_1 is then obtained from equation (4) by multiplying the electric field E_1 (*total*) by Q_1. In Cartesian coordinates, this force, expressed in newtons, is given by its components along the x and y axes by

$$F_1\ (total) = -1.44\hat{x} - 2.88\hat{y}.$$

The resulting force on Q_1 is in the direction of the total electric field at Q_1, shown in Figure 6. The magnitude of the force, which is obtained as the square root of the sum of the squares of the components of the force given in the above equation, equals 3.22 newtons.

Superposition principle

This calculation demonstrates an important property of the electromagnetic field known as the superposition principle. According to this principle, a field arising from a number of sources is determined by adding the individual fields from each source. The principle is illustrated by Figure 6, in which an electric field arising from several sources is determined by the superposition of the fields from each of the sources. In this case, the electric field at the location of Q_1 is the sum of the fields due to Q_2 and Q_3. Studies of electric fields over an extremely wide range of magnitudes have established the validity of the superposition principle.

By courtesy of the Department of Physics and Astronomy, Michigan State University

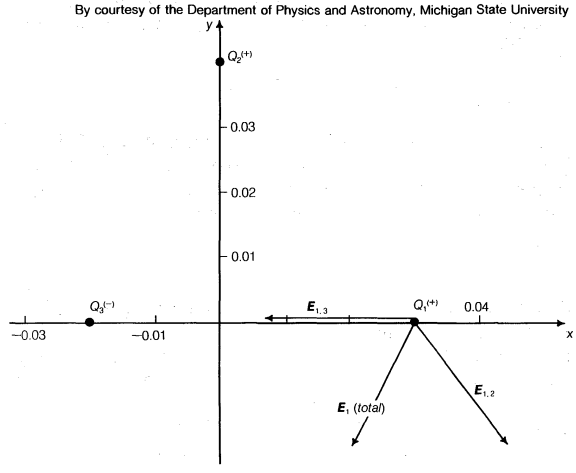


Figure 6: Electric field at the location of Q_1 (see text).

The vector nature of an electric field produced by a set of charges introduces a significant complexity. Specifying the field at each point in space requires giving both the magnitude and the direction at each location. In the Cartesian coordinate system, this necessitates knowing the magnitude of the x, y, and z components of the electric field at each point in space. It would be much simpler if the value of the electric field vector at any point in space could be derived from a scalar function with magnitude and sign.

The electric potential is just such a scalar function. Electric potential is related to the work done by an external force when it transports a charge slowly from one position to another in an environment containing other charges at rest. The difference between the potential at point A and the potential at point B is defined by the equation

$$V_A - V_B = \frac{\text{work to move charge } q \text{ from B to A}}{q}. \qquad (5)$$

As noted above, electric potential is measured in volts. Since work is measured in joules in the Système Internationale d'Unités (SI), one volt is equivalent to one joule per coulomb. The charge q is taken as a small test charge; it is assumed that the test charge does not disturb the distribution of the remaining charges during its transport from point B to point A.

To illustrate the work in equation (5), Figure 7 shows a positive charge $+Q$. Consider the work involved in moving a second charge q from B to A. Along path 1, work is done to offset the electric repulsion between the two charges. If path 2 is chosen instead, no work is done in moving q from B to C, since the motion is perpendicular to the electric force; moving q from C to D, the work is, by symmetry, identical as from B to A, and no work is required from D to A. Thus, the total work done in moving q from B to A is the same for either path. It can be shown easily that the same is true for any path going from B to A. When the initial and final positions of the charge q are located on a sphere centred on the location of the $+Q$ charge, no work is done; the electric potential at the initial position has the same value as at the final position. The

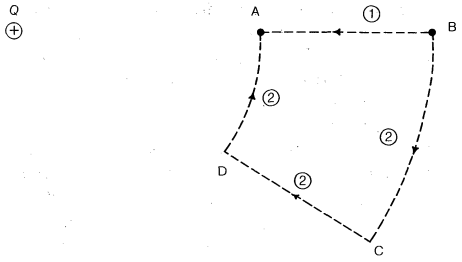


Figure 7: Positive charge +Q and two paths in moving a second charge, q, from B to A (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

sphere in this example is called an equipotential surface. When equation (5), which defines the potential difference between two points, is combined with Coulomb's law, it yields the following expression for the potential difference $V_A - V_B$ between points A and B:

$$V_A - V_B = k\frac{Q}{r_a} - k\frac{Q}{r_b}, \qquad (6)$$

where r_a and r_b are the distances of points A and B from Q. Choosing B far away from the charge Q and arbitrarily setting the electric potential to be zero far from the charge results in a simple equation for the potential at A:

$$V_A = k\frac{Q}{r_a}. \qquad (7)$$

The contribution of a charge to the electric potential at some point in space is thus a scalar quantity directly proportional to the magnitude of the charge and inversely proportional to the distance between the point and the charge. For more than one charge, one simply adds the contributions of the various charges. The result is a topological map that gives a value of the electric potential for every point in space.

Figure 8 provides three-dimensional views illustrating the effect of the positive charge $+Q$ located at the origin on either a second positive charge q (Figure 8A) or on a negative charge $-q$ (Figure 8B); the potential energy "landscape" is illustrated in each case. The potential energy of a charge q is the product qV of the charge and of the electric potential at the position of the charge. In Figure 8A, the positive charge q would have to be pushed by some external agent in order to get close to the location of $+Q$ because, as q approaches, it is subjected to an increasingly repulsive electric force. For the negative charge $-q$, the potential energy in Figure 8B shows, instead of a steep hill, a deep funnel. The electric potential due to $+Q$ is still positive, but the potential energy is negative, and the negative charge $-q$, in a manner quite analogous to a particle under the influence of gravity, is attracted toward the origin where charge $+Q$ is located.

The electric field is related to the variation of the electric potential in space. The potential provides a convenient tool for solving a wide variety of problems in electrostatics. In a region of space where the potential varies, a charge is subjected to an electric force. For a positive charge the direction of this force is opposite the gradient of the potential—that is to say, in the direction in which the potential decreases the most rapidly. A negative charge would be subjected to a force in the direction of the most rapid increase of the potential. In both instances, the magnitude of the force is proportional to the rate of change of the potential in the indicated directions. If the potential in a region of space is constant, there is no force on either positive or negative charge. In a 12-volt car battery, positive charges would tend to move away from the positive terminal and toward the negative terminal, while negative charges would tend to move in the opposite direction—*i.e.,* from the negative to the positive terminal. The latter occurs when a copper wire, in which there are electrons that are free to move, is connected between the two terminals of the battery.

The electric field has already been described in terms of the force on a charge. If the electric potential is known at every point in a region of space, the electric field can be derived from the potential. In vector calculus notation, the electric field is given by the negative of the gradient of the electric potential, $\boldsymbol{E} = -\mathbf{grad}\ V$. This expression specifies how the electric field is calculated at a given point. Since the field is a vector, it has both a direction and magnitude. The direction is that in which the potential decreases most rapidly, moving away from the point. The magnitude of the field is the change in potential across a small distance in the indicated direction divided by that distance.

Deriving electric field from potential

To become more familiar with the electric potential, a numerically determined solution is presented for a two-dimensional configuration of electrodes. A long, circular conducting rod is maintained at an electric potential of −20 volts. Next to the rod, a long L-shaped bracket, also made of conducting material, is maintained at a potential of +20 volts. Both the rod and bracket are placed inside a long, hollow metal tube with a square cross section; this enclosure is at a potential of zero (*i.e.,* it is at "ground" potential). Figure 9 shows the geometry of the problem. Because the situation is static, there is no electric field inside the material of the conductors. If there were such a field, the charges that are free to move in a conducting material would do so until equilibrium was reached. The

By courtesy of the Department of Physics and Astronomy, Michigan State University

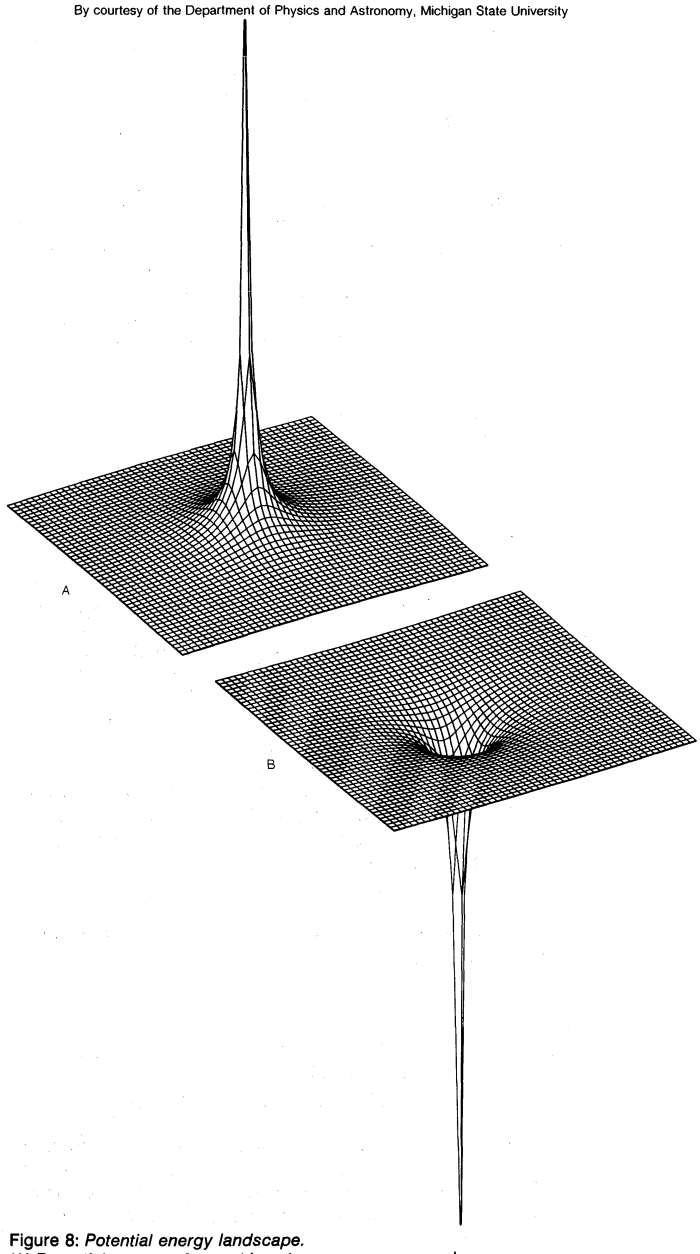


Figure 8: *Potential energy landscape.*
(A) Potential energy of a positive charge near a second positive charge. (B) Potential energy of a negative charge near a positive charge (see text).

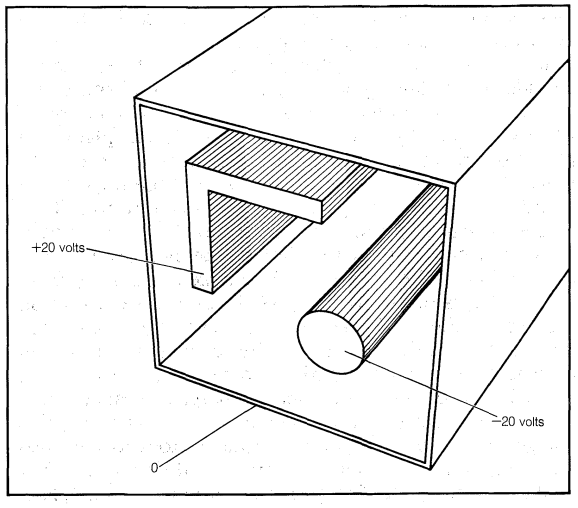


Figure 9: *Electrode configuration.*
A circular conducting rod is maintained at a potential of −20 volts, while an L-shaped bracket of conducting material is maintained at a potential of +20 volts. The electrodes are enclosed in a metal tube, which is at a potential of zero (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

charges are arranged so that their individual contributions to the electric field at points inside the conducting material add up to zero. In a situation of static equilibrium, excess charges are located on the surface of conductors. Because there are no electric fields inside the conducting material, all parts of a given conductor are at the same potential; hence, a conductor is an equipotential in a static situation.

In Figure 10, the numerical solution of the problem gives the potential at a large number of points inside the cavity. The locations of the +20-volt and −20-volt electrodes can be recognized easily. In carrying out the numerical solution of the electrostatic problem in the figure, the electrostatic potential was determined directly by means of one of its important properties: in a region where there is no charge (in this case, between the conductors), the value of the potential at a given point is the average of the values of the potential in the neighbourhood of the point. This follows from the fact that the electrostatic potential in a charge-free region obeys Laplace's equation, which in vector calculus notation is *div* **grad** $V = 0$. This equation is a special case of Poisson's equation *div* **grad** $V = \rho$, which is applicable to electrostatic problems in regions where the volume charge density is ρ. Laplace's

By courtesy of the Department of Physics and Astronomy, Michigan State University

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	4	8	9	10	10	10	10	8	5	3	1	1	0
0	8	20	20	20	20	20	20	20	9	5	2	1	0
0	9	20	20	20	20	20	20	20	9	4	2	1	0
0	10	20	20	18	16	14	12	9	5	2	0	0	0
0	10	20	20	16	12	9	5	1	−1	−2	−2	−1	0
0	10	20	20	14	9	3	−3	−7	−9	−8	−5	−3	0
0	10	20	20	12	5	−3	−11	−20	−20	−14	−9	−4	0
0	8	20	20	9	1	−7	−20	−20	−20	−20	−11	−5	0
0	5	9	9	5	−1	−9	−20	−20	−20	−20	−11	−5	0
0	3	5	4	2	−2	−8	−14	−20	−20	−14	−9	−4	0
0	1	2	2	0	−2	−5	−9	−11	−11	−9	−6	−3	0
0	1	1	1	0	−1	−3	−4	−5	−5	−4	−3	−2	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10: Numerical solution for the electrode configuration shown in Figure 9. The electrostatic potentials are in volts (see text).

equation states that the divergence of the gradient of the potential is zero in regions of space with no charge. In the example of Figure 10, the potential on the conductors remains constant. Arbitrary values of potential are initially assigned elsewhere inside the cavity. To obtain a solution, a computer replaces the potential at each coordinate point that is not on a conductor by the average of the values of the potential around that point; it scans the entire set of points many times until the values of the potentials differ by an amount small enough to indicate a satisfactory solution. Clearly, the larger the number of points, the more accurate the solution will be. The computation time as well as the computer memory size requirement increase rapidly, however, especially in three-dimensional problems with complex geometry. This method of solution is called the "relaxation" method.

In Figure 11, points with the same value of electric potential have been connected to reveal a number of important properties associated with conductors in static situations. The lines in the figure represent equipotential surfaces. The distance between two equipotential surfaces tells how rapidly the potential changes, with the smallest distances corresponding to the location of the greatest rate of change and thus to the largest values of the electric field. Looking at the +20-volt and +15-volt equipotential surfaces, one observes immediately that they are closest to each other at the sharp external corners of the right-

By courtesy of the Department of Physics and Astronomy, Michigan State University

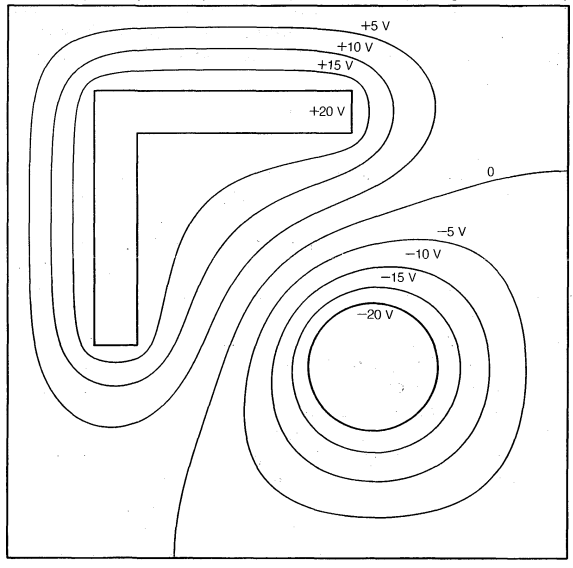


Figure 11: *Equipotential surfaces.*
The distance between two equipotential surfaces, represented by the lines, indicates how rapidly the potential changes. The smallest distances correspond to the location of the greatest rate of change and therefore to the largest values of the electric field.

angle conductor. This shows that the strongest electric fields on the surface of a charged conductor are found on the sharpest external parts of the conductor; electrical breakdowns are most likely to occur there. It also should be noted that the electric field is weakest in the inside corners, both on the inside corner of the right-angle piece and on the inside corners of the square enclosure.

In Figure 12, dashed lines indicate the direction of the electric field. The strength of the field is reflected by the density of these dashed lines. Again, it can be seen that the field is strongest on outside corners of the charged L-shaped conductor; the largest surface charge density must occur at those locations. The field is weakest in the inside corners. The signs of the charges on the conducting surfaces can be deduced from the fact that electric fields point away from positive charges and toward negative charges. The magnitude of the surface charge density σ on the conductors is measured in coulombs per metre squared and is given by

$$\sigma = \varepsilon_0 E, \qquad (8)$$

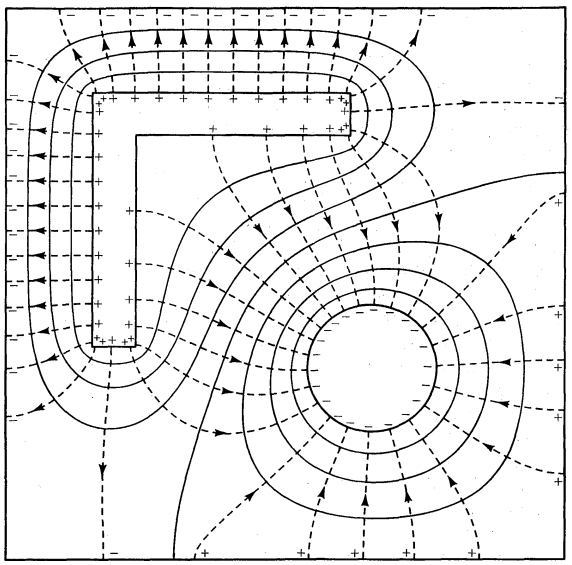


Figure 12: *Electric field lines.*
The density of the dashed lines indicates the strength of the field (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

where ε_0 is called the permittivity of free space and has the value of 8.854×10^{-12} coulomb squared per newton-square metre. In addition, ε_0 is related to the constant k in Coulomb's law by

$$k = \frac{1}{4\pi\varepsilon_0}. \tag{9}$$

Field lines and equipotential surfaces

Figure 12 also illustrates an important property of an electric field in static situations: field lines are always perpendicular to equipotential surfaces. The field lines meet the surfaces of the conductors at right angles, since these surfaces also are equipotentials. Figure 13 completes this example by showing the potential energy landscape of a small positive charge q in the region. From the variation in potential energy, it is easy to picture how electric forces tend to drive the positive charge q from higher to lower potential—*i.e.*, from the L-shaped bracket at +20 volts toward the square-shaped enclosure at ground (0 volts) or toward the cylindrical rod maintained at a potential of −20 volts. It also graphically displays the strength of force near the sharp corners of conducting electrodes.

Capacitance. A useful device for storing electrical energy consists of two conductors in close proximity and insulated from each other. A simple example of such a storage device is the parallel-plate capacitor. If positive charges with total charge $+Q$ are deposited on one of the conductors and an equal amount of negative charge $-Q$ is deposited on the second conductor, the capacitor is said to have a charge Q. As shown in Figure 14, it consists of two flat conducting plates, each of area A, parallel to each other and separated by a distance d.

By courtesy of the Department of Physics and Astronomy, Michigan State University

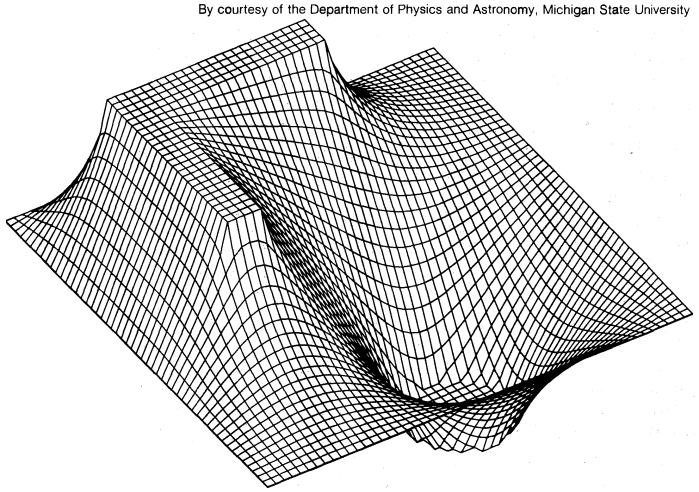


Figure 13: Potential energy for a positive charge (see text).

Principle of the capacitor

To understand how a charged capacitor stores energy, consider the following charging process. With both plates of the capacitor initially uncharged, a small amount of negative charge is removed from the lower plate and placed on the upper plate. Thus, little work is required to make the lower plate slightly positive and the upper plate slightly negative. As the process is repeated, however, it becomes increasingly difficult to transport the same amount of negative charge, since the charge is being moved toward a plate that is already negatively charged and away from a plate that is positively charged. The negative charge on the upper plate repels the negative charge moving toward it, and the positive charge on the lower plate exerts an attractive force on the negative charge being moved away. Therefore, work has to be done to charge the capacitor.

Where and how is this energy stored? The negative charges on the upper plate are attracted toward the positive charges on the lower plate and could do work if they could leave the plate. Because they cannot leave the plate, however, the energy is stored. A mechanical analogy is the potential energy of a stretched spring. Another way to understand the energy stored in a capacitor is to compare an uncharged capacitor with a charged capacitor. In the uncharged capacitor, there is no electric field between the plates; in the charged capacitor, because of the positive and negative charges on the inside surfaces of the plates, there is an electric field between the plates with the field lines pointing from the positively charged plate to the negatively charged one. The energy stored is the energy that

By courtesy of the Department of Physics and Astronomy, Michigan State University

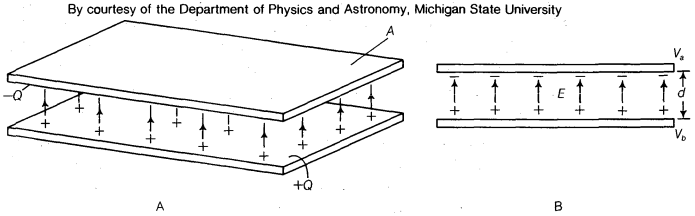


Figure 14: *Parallel-plate capacitor.*
(A) This storage device consists of two flat conducting plates, each of area *A*. (B) These plates are parallel and separated by a small distance *d* (see text).

was required to establish the field. In the simple geometry of Figure 14, it is apparent that there is a nearly uniform electric field between the plates; the field becomes more uniform as the distance between the plates decreases and the area of the plates increases. It was explained above how the magnitude of the electric field can be obtained from the electric potential. In summary, the electric field is the change in the potential across a small distance in a direction perpendicular to an equipotential surface divided by that small distance. In Figure 14, the upper plate is assumed to be at a potential of V_a volts, and the lower plate at a potential of V_b volts. The size of the electric field is

$$E = \frac{V_b - V_a}{d} \tag{10}$$

in volts per metre, where d is the separation of the plates. If the charged capacitor has a total charge of $+Q$ on the inside surface of the lower plate (it is on the inside surface because it is attracted to the negative charges on the upper plate), the positive charge will be uniformly distributed on the surface with the value

$$\sigma = \frac{Q}{A} \tag{11}$$

in coulombs per metre squared. Equation (8) gives the electric field when the surface charge density is known as $E = \sigma/\varepsilon_0$. This, in turn, relates the potential difference to the charge on the capacitor and the geometry of the plates. The result is

$$V_b - V_a = \frac{Qd}{\varepsilon_0 A} = \frac{Q}{C}. \tag{12}$$

The quantity C is termed capacity; for the parallel-plate capacitor, C is equal to $\varepsilon_0 A/d$. The unit used for capacity

is the farad (F); one farad equals one coulomb per volt. In equation (12), only the potential difference is involved. The potential of either plate can be set arbitrarily without altering the electric field between the plates. Often one of the plates is grounded—*i.e.*, its potential is set at the Earth potential, which is referred to as zero volts. The potential difference is then denoted as ΔV, or simply as V.

Three equivalent formulas for the total energy W of a capacitor with charge Q and potential difference V are

$$W=\frac{1}{2}\frac{Q^2}{C}=\frac{1}{2}CV^2=\frac{1}{2}QV. \quad (13)$$

All are expressed in joules. The stored energy in the parallel-plate capacitor also can be expressed in terms of the electric field; it is, in joules,

$$W=\frac{1}{2}\varepsilon_0 E^2\,(Ad). \quad (14)$$

The quantity Ad, the area of each plate times the separation of the two plates, is the volume between the plates. Thus, the energy per unit volume (*i.e.*, the energy density of the electric field) is given by $\frac{1}{2}\varepsilon_0 E^2$ in units of joules per metre cubed.

The amount of charge stored in a capacitor is the product of the voltage and the capacity. What limits the amount of charge that can be stored on a capacitor? The voltage can be increased, but electric breakdown will occur if the electric field inside the capacitor becomes too large. The capacity can be increased by expanding the electrode areas and by reducing the gap between the electrodes. In general, capacitors that can withstand high voltages have a relatively small capacity. If only low voltages are needed, however, compact capacitors with rather large capacities can be manufactured. One method for increasing capacity is to insert between the conductors an insulating material that reduces the voltage because of its effect on the electric field. Such materials are called dielectrics (substances with no free charges). When the molecules of a dielectric are placed in the electric field, their negatively charged electrons separate slightly from their positively charged cores. With this separation, referred to as polarization, the molecules acquire an electric dipole moment. A cluster of charges with an electric dipole moment is often called an electric dipole.

Polarization and electric dipole moment

Is there an electric force between a charged object and uncharged matter, such as a piece of wood? Surprisingly, the answer is yes, and the force is attractive. The reason is that under the influence of the electric field of a charged object, the negatively charged electrons and positively charged nuclei within the atoms and molecules are subjected to forces in opposite directions. As a result, the negative and positive charges separate slightly. Such atoms and molecules are said to be polarized and to have an electric dipole moment. The molecules in the wood acquire an electric dipole moment in the direction of the external electric field. The polarized molecules are attracted toward the charged object because the field increases in the direction of the charged object.

The electric dipole moment $\boldsymbol{p}$ of two charges $+q$ and $-q$ separated by a distance l is a vector of magnitude $p=ql$ with a direction from the negative to the positive charge. An electric dipole in an external electric field is subjected to a torque $\tau=pE\sin\theta$, where θ is the angle between $\boldsymbol{p}$ and $\boldsymbol{E}$. The torque tends to align the dipole moment $\boldsymbol{p}$ in the direction of $\boldsymbol{E}$. The potential energy of the dipole is given by $U_e=-pE\cos\theta$, or in vector notation $U_e=-\boldsymbol{p}\cdot\boldsymbol{E}$. In a nonuniform electric field, the potential energy of an electric dipole also varies with position, and the dipole can be subjected to a force. The force on the dipole is in the direction of increasing field when $\boldsymbol{p}$ is aligned with $\boldsymbol{E}$, since the potential energy U_e decreases in that direction.

The polarization of a medium $\boldsymbol{P}$ gives the electric dipole moment per unit volume of the material; it is expressed in units of coulombs per metre squared. When a dielectric is placed in an electric field, it acquires a polarization that depends on the field. The electric susceptibility χ_e relates the polarization to the electric field as $\boldsymbol{P}=\chi_e\boldsymbol{E}$. In general, χ_e varies slightly depending on the strength of the electric field, but for some materials, called linear dielectrics, it is a constant. The dielectric constant κ of a substance is related to its susceptibility as $\kappa=1+\chi_e/\varepsilon_0$; it is a dimensionless quantity. Table 1 lists the dielectric constants of a few substances.

Table 1: Dielectric Constants of Some Materials (at room temperature)

material	dielectric constant (κ)
Vacuum	1.0
Air	1.0006
Oil	2.2
Polyethylene	2.26
Beeswax	2.8
Fused quartz	3.78
Water	80
Calcium titanate	168
Barium titanate	1,250

The presence of a dielectric affects many electric quantities. A dielectric reduces by a factor K the value of the electric field and consequently also the value of the electric potential from a charge within the medium. As seen in Table 1, a dielectric can have a large effect. The insertion of a dielectric between the electrodes of a capacitor with a given charge reduces the potential difference between the electrodes and thus increases the capacitance of the capacitor by the factor K. For a parallel-plate capacitor filled with a dielectric, the capacity becomes $C=K\varepsilon_0 A/d$. A third and important effect of a dielectric is to reduce the speed of electromagnetic waves in a medium by the factor $\sqrt{K}$.

Effects of dielectrics

Capacitors come in a wide variety of shapes and sizes. Not all have parallel plates; some are cylinders, for example. If two plates, each one square centimetre in area, are separated by a dielectric with $K=2$ of one millimetre thickness, the capacity is 1.76×10^{-12} F, about two picofarads. Charged to 20 volts, this capacitor would store about 40 picocoulombs of charge; the electric energy stored would be 400 picojoules. Even small-sized capacitors can store enormous amounts of charge. Modern techniques and dielectric materials permit the manufacture of capacitors that occupy less than one cubic centimetre and yet store 10^{10} times more charge and electric energy than in the above example.

Capacitors have many important applications. They are used, for example, in digital circuits so that information stored in large computer memories is not lost during a momentary electric power failure; the electric energy stored in such capacitors maintains the information during the temporary loss of power. Capacitors play an even more important role as filters to divert spurious electric signals and thereby prevent damage to sensitive components and circuits caused by electric surges. How capacitors provide such protection is discussed below in the section *Transient response.*

DIRECT ELECTRIC CURRENT

Basic phenomena and principles. Many electric phenomena occur under what is termed steady-state conditions. This means that such electric quantities as current, voltage, and charge distributions are not affected by the passage of time. For instance, because the current through a filament inside a car headlight does not change with time, the brightness of the headlight remains constant. An example of a nonsteady-state situation is the flow of charge between two conductors that are connected by a thin conducting wire and that initially have an equal but opposite charge. As current flows from the positively charged conductor to the negatively charged one, the charges on both conductors decrease with time, as does the potential difference between the conductors. The current therefore also decreases with time and eventually ceases when the conductors are discharged.

Steady-state phenomena

In an electric circuit under steady-state conditions, the flow of charge does not change with time and the charge distribution stays the same. Since charge flows from one location to another, there must be some mechanism to keep the charge distribution constant. In turn, the values

of the electric potentials remain unaltered with time. Any device capable of keeping the potentials of electrodes unchanged as charge flows from one electrode to another is called a source of electromotive force, or simply an emf.

Figure 15 shows a wire made of a conducting material such as copper. By some external means, an electric field is established inside the wire in a direction along its length.

By courtesy of the Department of Physics and Astronomy, Michigan State University

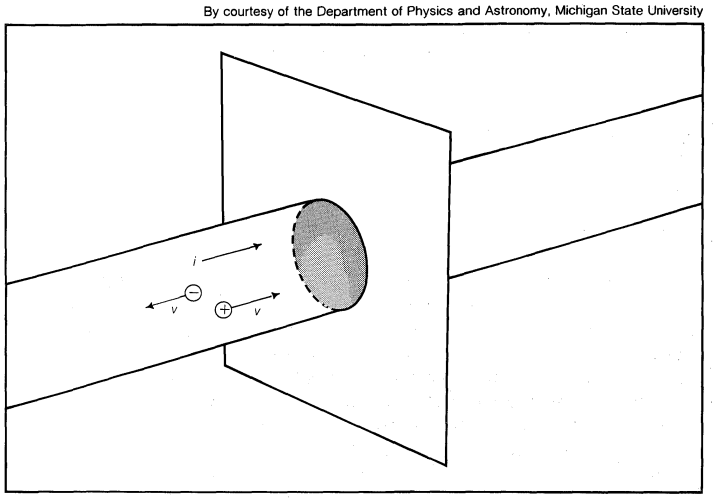


Figure 15: Motion of charge in electric current i (see text).

The electrons that are free to move will gain some speed. Since they have a negative charge, they move in the direction opposite that of the electric field. The current i is defined to have a positive value in the direction of flow of positive charges. If the moving charges that constitute the current i in a wire are electrons, the current is a positive number when it is in a direction opposite to the motion of the negatively charged electrons. (If the direction of motion of the electrons were also chosen to be the direction of a current, the current would have a negative value.) The current is the amount of charge crossing a plane transverse to the wire per unit time—*i.e.*, in a period of one second. If there are n free particles of charge q per unit volume with average velocity v and the cross-sectional area of the wire is A, the current i, in elementary calculus notation, is

$$i = \frac{dQ}{dt} = nevA, \tag{15}$$

where dQ is the amount of charge that crosses the plane in a time interval dt. The unit of current is the ampere (A); one ampere equals one coulomb per second. A useful quantity related to the flow of charge is current density, the flow of current per unit area. Symbolized by $\boldsymbol{J}$, it has a magnitude of i/A and is measured in amperes per square metre.

Wires of different materials have different current densities for a given value of the electric field $\boldsymbol{E}$; for many materials, the current density is directly proportional to the electric field. This behaviour is represented by Ohm's law:

$$\boldsymbol{J} = \sigma_J \boldsymbol{E}. \tag{16}$$

The proportionality constant σ_J is the conductivity of the material. In a metallic conductor, the charge carriers are electrons and, under the influence of an external electric field, they acquire some average drift velocity in the direction opposite the field. In conductors of this variety, the drift velocity is limited by collisions, which heat the conductor.

If the wire in Figure 15 has a length l and area A and if an electric potential difference of V is maintained between the ends of the wire, a current i will flow in the wire. The electric field $\boldsymbol{E}$ in the wire has a magnitude V/l. The equation for the current, using Ohm's law, is

$$i = JA = \frac{\sigma_J V}{l} A \tag{17}$$

or

$$V = i\frac{l}{\sigma_J A}. \tag{18}$$

The quantity $l/\sigma_J A$, which depends on both the shape and material of the wire, is called the resistance R of the wire. Resistance is measured in ohms (Ω). The equation for resistance,

$$R = \frac{l}{\sigma_J A}, \tag{19}$$

is often written as

$$R = \frac{\rho l}{A}, \tag{20}$$

where ρ is the resistivity of the material and is simply $1/\sigma_J$. The geometric aspects of resistance in equation (20) are easy to appreciate: the longer the wire, the greater the resistance to the flow of charge. A greater cross-sectional area results in a smaller resistance to the flow. Resistivity

The resistive strain gauge is an important application of equation (20). Strain, $\delta l/l$, is the fractional change in the length of a body under stress, where δl is the change of length and l is the length. The strain gauge consists of a thin wire or narrow strip of a metallic conductor such as constantan, an alloy of nickel and copper. A strain changes the resistance because the length, area, and resistivity of the conductor change. In constantan, the fractional change in resistance $\delta R/R$ is directly proportional to the strain with a proportionality constant of approximately 2.

A common form of Ohm's law is Ohm's law

$$V = iR, \tag{21}$$

where V is the potential difference in volts between the two ends of an element with an electric resistance of R ohms and where i is the current through that element.

Table 2 lists the resistivities of certain materials at room temperature. These values depend to some extent on temperature; therefore, in applications where the temperature is very different from room temperature, the proper values of resistivities must be used to calculate the resistance. As an example, equation (20) shows that a copper wire 59 metres long and with a cross-sectional area of one square millimetre has an electric resistance of one ohm at room temperature.

Table 2: Electric Resistivities
(at room temperature)

material	resistivity ρ (ohm metre)
Silver	$1.6 \cdot 10^{-8}$
Copper	$1.7 \cdot 10^{-8}$
Aluminum	$2.7 \cdot 10^{-8}$
Carbon (graphite)	$1.4 \cdot 10^{-5}$
Germanium*	$4.7 \cdot 10^{-1}$
Silicon*	$2 \cdot 10^{3}$
Carbon (diamond)	$5 \cdot 10^{12}$
Polyethylene	$1 \cdot 10^{17}$
Fused quartz	$> 1 \cdot 10^{19}$

*Values very sensitive to purity.

Conductors, insulators, and semiconductors. Materials are classified as conductors, insulators, or semiconductors according to their electric conductivity. The classifications can be understood in atomic terms. Electrons in an atom can have only certain well-defined energies, and, depending on their energies, the electrons are said to occupy particular energy levels. In a typical atom with many electrons, the lower energy levels are filled, each with the number of electrons allowed by a quantum mechanical rule known as the Pauli exclusion principle. Depending on the element, the highest energy level to have electrons may or may not be completely full. If two atoms of some element are brought close enough together so that they interact, the two-atom system has two closely spaced levels for each level of the single atom. If 10 atoms interact, the 10-atom system will have a cluster of 10 levels corresponding to each single level of an individual atom. In a solid, the number of atoms and hence the number of levels is extremely large; most of the higher energy levels overlap in a continuous fashion except for certain energies in which there are no levels at all. Energy regions with levels are called energy bands, and regions that have no levels are referred to as band gaps. Energy bands and band gaps

The highest energy band occupied by electrons is the valence band. In a conductor, the valence band is partially filled, and since there are numerous empty levels, the electrons are free to move under the influence of an electric field; thus, in a metal the valence band is also the conduction band. In an insulator, electrons completely fill the valence band; and the gap between it and the next band, which is the conduction band, is large. The electrons cannot move under the influence of an electric field unless they are given enough energy to cross the large energy gap to the conduction band. In a semiconductor, the gap to the conduction band is smaller than in an insulator. At room temperature, the valence band is almost completely filled. A few electrons are missing from the valence band because they have acquired enough thermal energy to cross the band gap to the conduction band; as a result, they can move under the influence of an external electric field. The "holes" left behind in the valence band are mobile charge carriers but behave like positive charge carriers.

For many materials, including metals, resistance to the flow of charge tends to increase with temperature. For example, an increase of 5° C (9° F) increases the resistivity of copper by 2 percent. In contrast, the resistivity of insulators and especially of semiconductors such as silicon and germanium decreases rapidly with temperature; the increased thermal energy causes some of the electrons to populate levels in the conduction band where, influenced by an external electric field, they are free to move. The energy difference between the valence levels and the conduction band has a strong influence on the conductivity of these materials, with a smaller gap resulting in higher conduction at lower temperatures.

Range of resistivities in different materials

The values of electric resistivities listed in Table 2 show an extremely large variation in the capability of different materials to conduct electricity. The principal reason for the large variation is the wide range in the availability and mobility of charge carriers within the materials. The copper wire in Figure 15, for example, has many extremely mobile carriers; each copper atom has approximately one free electron, which is highly mobile because of its small mass. An electrolyte, such as a saltwater solution, is not as good a conductor as copper. The sodium and chlorine ions in the solution provide the charge carriers. The large mass of each sodium and chlorine ion increases as other attracted ions cluster around them. As a result, the sodium and chlorine ions are far more difficult to move than the free electrons in copper. Pure water also is a conductor, although it is a poor one because only a very small fraction of the water molecules are dissociated into ions. The oxygen, nitrogen, and argon gases that make up the atmosphere are somewhat conductive because a few charge carriers form when the gases are ionized by radiation from radioactive elements on the Earth as well as from extraterrestrial cosmic rays (*i.e.*, high-speed atomic nuclei and electrons). Electrophoresis is an interesting application based on the mobility of particles suspended in an electrolytic solution. Different particles (proteins, for example) move in the same electric field at different speeds; the difference in speed can be utilized to separate the contents of the suspension.

A current flowing through a wire heats it. This familiar phenomenon occurs in the heating coils of an electric range or in the hot tungsten filament of an electric light bulb. This ohmic heating is the basis for the fuses used to protect electric circuits and prevent fires; if the current exceeds a certain value, a fuse, which is made of an alloy with a low melting point, melts and interrupts the flow of current. The power P dissipated in a resistance R through which current i flows is given by

$$P = i^2R, \tag{22}$$

where P is in watts (one watt equals one joule per second), i is in amperes, and R is in ohms. According to Ohm's law, the potential difference V between the two ends of the resistor is given by $V = iR$, and so the power P can be expressed equivalently as

$$P = iV = \frac{V^2}{R}. \tag{23}$$

Superconductivity

In certain materials, however, the power dissipation that manifests itself as heat suddenly disappears if the conductor is cooled to a very low temperature. The disappearance of all resistance is a phenomenon known as superconductivity. As mentioned earlier, electrons acquire some average drift velocity v under the influence of an electric field in a wire. Normally the electrons, subjected to a force because of an electric field, accelerate and progressively acquire greater speed. Their velocity is, however, limited in a wire because they lose some of their acquired energy to the wire in collisions with other electrons and in collisions with atoms in the wire. The lost energy is either transferred to other electrons, which later radiate, or the wire becomes excited with tiny mechanical vibrations referred to as phonons. Both processes heat the material. The term phonon emphasizes the relationship of these vibrations to another mechanical vibration—namely, sound. In a superconductor, a complex quantum mechanical effect prevents these small losses of energy to the medium. The effect involves interactions between electrons and also those between electrons and the rest of the material. It can be visualized by considering the coupling of the electrons in pairs with opposite momenta; the motion of the paired electrons is such that no energy is given up to the medium in inelastic collisions or phonon excitations. One can imagine that an electron about to "collide" with and lose energy to the medium could end up instead colliding with its partner so that they exchange momentum without imparting any to the medium.

A superconducting material widely used in the construction of electromagnets is an alloy of niobium and titanium. This material must be cooled to a few degrees above absolute zero temperature, −263.66° C (or 9.5 K), in order to exhibit the superconducting property. Such cooling requires the use of liquefied helium, which is rather costly. During the late 1980s, materials that exhibit superconducting properties at much higher temperatures were discovered. These temperatures are higher than the −196° C of liquid nitrogen, making it possible to use the latter instead of liquid helium. Since liquid nitrogen is plentiful and cheap, such materials may provide great benefits in a wide variety of applications, ranging from electric power transmission to high-speed computing.

Electromotive force. A 12-volt automobile battery can deliver current to a circuit such as that of a car radio for a considerable length of time, during which the potential difference between the terminals of the battery remains close to 12 volts. The battery must have a means of continuously replenishing the excess positive and negative charges that are located on the respective terminals and that are responsible for the 12-volt potential difference between the terminals. The charges must be transported from one terminal to the other in a direction opposite to the electric force on the charges between the terminals. Any device that accomplishes this transport of charge constitutes a source of electromotive force. A car battery, for example, uses chemical reactions to generate electromotive force. The Van de Graaff generator shown in Figure 16 is a mechanical device that produces an electromotive force. Invented by the American physicist Robert J. Van de Graaff in the 1930s, this type of particle accelerator has been widely used to study subatomic particles. Because it is conceptually simpler than a chemical source of electromotive force, the Van de Graaff generator will be discussed first.

Sources of electromotive force

An insulating conveyor belt carries positive charge from the base of the Van de Graaff machine to the inside of a large conducting dome. The charge is removed from the belt by the proximity of sharp metal electrodes called charge remover points. The charge then moves rapidly to the outside of the conducting dome. The positively charged dome creates an electric field, which points away from the dome and provides a repelling action on additional positive charges transported on the belt toward the dome. Thus, work is done to keep the conveyor belt turning. If a current is allowed to flow from the dome to ground and if an equal current is provided by the transport of charge on the insulating belt, equilibrium is established and the potential of the dome remains at a constant positive value.

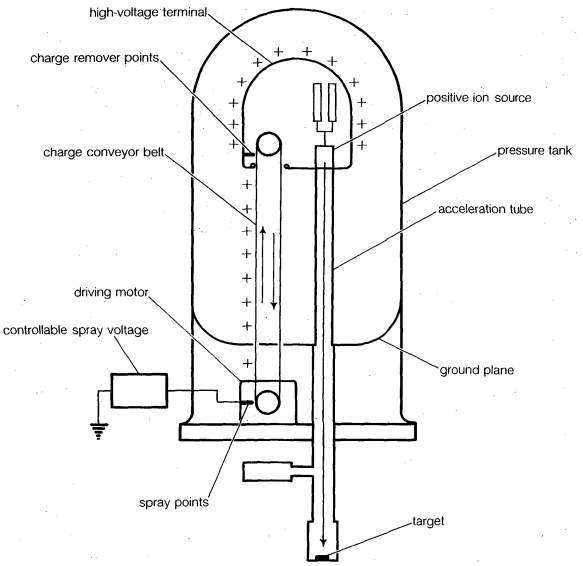


Figure 16: Van de Graaff accelerator.

From I. Kaplan, *Nuclear Physics*, © 1962, Addison-Wesley Publishing Co., Inc., Reading, Mass.; reprinted with permission of the publisher

In this example, the current from the dome to ground consists of a stream of positive ions inside the accelerating tube, moving in the direction of the electric field. The motion of the charge on the belt is in a direction opposite to the force that the electric field of the dome exerts on the charge. This motion of charge in a direction opposite the electric field is a feature common to all sources of electromotive force.

In the case of a chemically generated electromotive force, chemical reactions release energy. If these reactions take place with chemicals in close proximity to each other (*e.g.*, if they mix), the energy released heats the mixture. To produce a voltaic cell, these reactions must occur in separate locations. A copper wire and a zinc wire poked into a lemon make up a simple voltaic cell. The potential difference between the copper and the zinc wires can be measured easily and is found to be 1.1 volts; the copper wire acts as the positive terminal. Such a "lemon battery" is a rather poor voltaic cell capable of supplying only small amounts of electric power. Another kind of 1.1-volt battery constructed with essentially the same materials can provide much more electricity. In this case, a copper wire is placed in a solution of copper sulfate and a zinc wire in a solution of zinc sulfate; the two solutions are connected electrically by a potassium chloride salt bridge. (A salt bridge is a conductor with ions as charge carriers.) In both kinds of batteries, the energy comes from the difference in the degree of binding between the electrons in copper and those in zinc. Energy is gained when copper ions from the copper sulfate solution are deposited on the copper electrode as neutral copper ions, thus removing free electrons from the copper wire. At the same time, zinc atoms from the zinc wire go into solution as positively charged zinc ions, leaving the zinc wire with excess free electrons. The result is a positively charged copper wire and a negatively charged zinc wire. The two reactions are separated physically, with the salt bridge completing the internal circuit.

Figure 17 illustrates a 12-volt lead-acid battery, using standard symbols for depicting batteries in a circuit. The battery consists of six voltaic cells, each with an electromotive force of approximately two volts; the cells are connected in series, so that the six individual voltages add up to about 12 volts (Figure 17A). As shown in Figure 17B, each two-volt cell consists of a number of positive and negative electrodes connected electrically in parallel. The parallel connection is made to provide a large surface area of electrodes, on which chemical reactions can take place. The higher rate at which the materials of the electrodes are able to undergo chemical transformations allows the battery to deliver a larger current.

In the lead-acid battery, each voltaic cell consists of a negative electrode of pure, spongy lead (Pb) and a positive electrode of lead oxide (PbO_2). Both the lead and lead oxide are in a solution of sulfuric acid (H_2SO_4) and water (H_2O). At the positive electrode, the chemical reaction is $PbO_2 + SO_4^{--} + 4H^+ + 2e^- \rightarrow PbSO_4 + 2H_2O + (1.68\ V)$. At the negative terminal, the reaction is $Pb + SO_4^{--} \rightarrow PbSO_4 + 2e^- + (0.36\ V)$. The cell potential is $1.68 + 0.36 = 2.04$ volts. The 1.68 and 0.36 volts in the above equations are, respectively, the reduction and oxidation potentials; they are related to the binding of the electrons in the chemicals. When the battery is recharged, either by a car generator or by an external power source, the two chemical reactions are reversed.

Direct-current circuits. The simplest direct-current (DC) circuit consists of a resistor connected across a source of electromotive force. The symbol for a resistor is shown in Figure 18; here the value of R, 60Ω, is given by the numerical value adjacent to the symbol. The symbol for a source of electromotive force, E, is shown with the associated value of the voltage. Convention gives the terminal with the long line a higher (*i.e.*, more positive) potential than the terminal with the short line. Straight lines connecting various elements in a circuit are assumed to have negligible resistance, so that there is no change in potential across these connections. The circuit shows a 12-volt electromotive force connected to a 60Ω resistor. The letters *a*, *b*, *c*, and *d* on the diagram are reference points.

Circuit elements

The function of the source of electromotive force is to maintain point *a* at a potential 12 volts more positive than point *d*. Thus, the potential difference $V_a - V_d$ is 12 volts. The potential difference across the resistance is $V_b - V_c$. From Ohm's law, the current i flowing through the resistor is

$$i = \frac{V_b - V_c}{R} = \frac{V_b - V_c}{60}. \qquad (24)$$

Since points *a* and *b* are connected by a conductor of negligible resistance, they are at the same potential. For the same reason, *c* and *d* are at the same potential. Therefore, $V_b - V_c = V_a - V_d = 12$ volts. The current in the circuit is given by equation (24). Thus, $i = 12/60 = 0.2$ ampere. The power dissipated in the resistor as heat is easily calculated using equation (22):

$$P = i^2R = (0.02)^2 \times 60 = 2.4 \text{ watts}.$$

Where does the energy that is dissipated as heat in the resistor come from? It is provided by a source of elec-

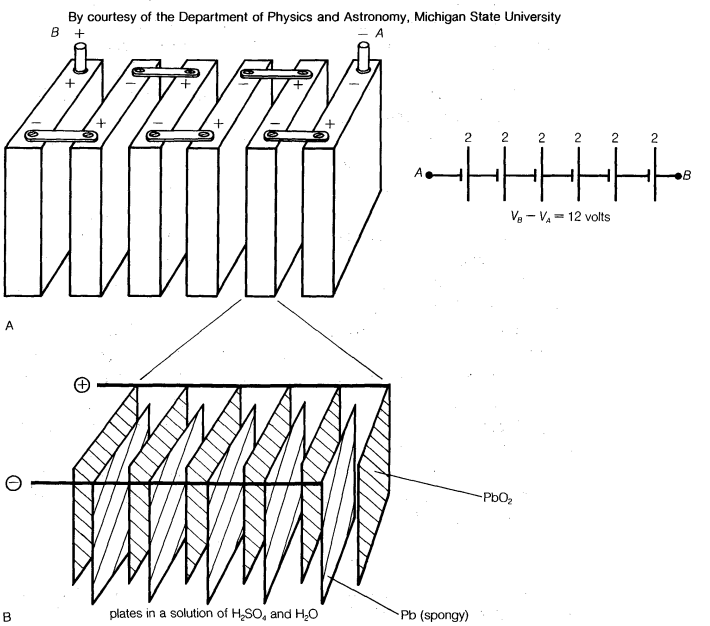


Figure 17: *Voltaic cells and electrodes of a 12-volt lead-acid battery.*
(A) The battery consists of six two-volt cells connected in series. (B) Each component cell is composed of several negative and positive electrodes made of pure spongy lead and lead oxide, respectively; the electrodes, connected in parallel, are immersed in a dilute solution of sulfuric acid.

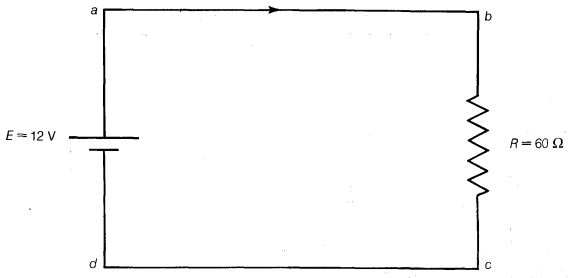


Figure 18: Direct-current circuit (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

tromotive force (*e.g.*, a lead-acid battery). Within such a source, for each amount of charge dQ moved from the lower potential at d to the higher potential at a, an amount of work is done equal to $dW = dQ(V_a - V_d)$. If this work is done in a time interval dt, the power delivered by the battery is obtained by dividing dW by dt. Thus, the power delivered by the battery (in watts) is

$$\frac{dW}{dt} = (V_a - V_d)\frac{dQ}{dt} = (V_a - V_d)i.$$

Using the values $i = 0.2$ ampere and $V_a - V_d = 12$ volts makes $dW/dt = 2.4$ watts. As expected, the power delivered by the battery is equal to the power dissipated as heat in the resistor.

Resistors in series and parallel. If two resistors are connected in Figure 19A so that all of the electric charge must traverse both resistors in succession, the equivalent resistance to the flow of current is the sum of the resistances.

By courtesy of the Department of Physics and Astronomy, Michigan State University

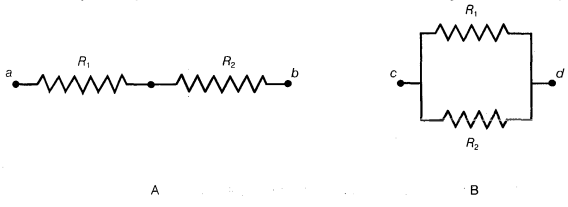


Figure 19: *Resistors.*
(A) In series. (B) In parallel.

Using R_1 and R_2 for the individual resistances, the resistance between a and b is given by

$$R_{ab} = R_1 + R_2. \qquad (25a)$$

This result can be appreciated by thinking of the two resistors as two pieces of the same type of thin wire. Connecting the wires in series as shown simply increases their length to equal the sum of their two lengths. As equation (20) indicates, the resistance is the same as that given by equation (25a). The resistances R_1 and R_2 can be replaced in a circuit by the equivalent resistance R_{ab}. If $R_1 = 5\Omega$ and $R_2 = 2\Omega$, then $R_{ab} = 7\Omega$. If two resistors are connected as shown in Figure 19B, the electric charges have alternate paths for flowing from c to d. The resistance to the flow of charge from c to d is clearly less than if either R_1 or R_2 were missing. Anyone who has ever had to find a way out of a crowded theatre can appreciate how much easier it is to leave a building with several exits than one with a single exit. The value of the equivalent resistance for two resistors in parallel is given by the equation

$$\frac{1}{R_{cd}} = \frac{1}{R_1} + \frac{1}{R_2}. \qquad (25b)$$

This relationship follows directly from the definition of resistance in equation (20), where $1/R$ is proportional to the area. If the resistors R_1 and R_2 are imagined to be wires of the same length and material, they would be wires with different cross-sectional areas. Connecting them in parallel is equivalent to placing them side by side, increasing the total area available for the flow of charge. Clearly, the equivalent resistance is smaller than the resistance of either resistor individually. As a numerical example, for $R_1 = 5\Omega$ and $R_2 = 2\Omega$, $1/R_{cd} = 1/5 + 1/2 = 0.7$. Therefore, $R_{cd} = 1/0.7 = 1.43\Omega$. As expected, the equivalent resistance of 1.43 ohms is smaller than either 2 ohms or 5 ohms. It should be noted that both equations (25a) and (25b) are given in a form in which they can be extended easily to any number of resistances.

Kirchhoff's laws of electric circuits. Two simple relationships can be used to determine the value of currents in circuits. They are useful even in rather complex situations such as circuits with multiple loops. The first relationship deals with currents at a junction of conductors. Figure 20 shows three such junctions, with the currents assumed to flow in the directions indicated.

Determining the value of currents in circuits

By courtesy of the Department of Physics and Astronomy, Michigan State University

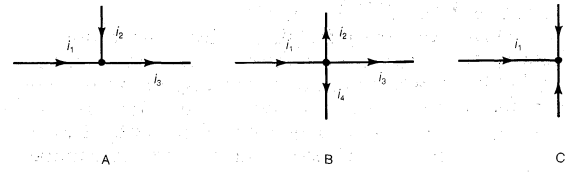


Figure 20: Electric currents at a junction (see text).

Simply stated, the sum of currents entering a junction equals the sum of currents leaving that junction. This statement is commonly called Kirchhoff's first law (after the German physicist Gustav Robert Kirchhoff, who formulated it). For Figure 20A, the sum is $i_1 + i_2 = i_3$. For Figure 20B, $i_1 = i_2 + i_3 + i_4$. For Figure 20C, $i_1 + i_2 + i_3 = 0$. If this last equation seems puzzling because all the currents appear to flow in and none flows out, it is because of the choice of directions for the individual currents. In solving a problem, the direction chosen for the currents is arbitrary. Once the problem has been solved, some currents have a positive value, and the direction arbitrarily chosen is the one of the actual current. In the solution some currents may have a negative value, in which case the actual current flows in a direction opposite that of the arbitrary initial choice.

Kirchhoff's second law is as follows: the sum of electromotive forces in a loop equals the sum of potential drops in the loop. When electromotive forces in a circuit are symbolized as circuit components as in Figure 18, this law can be stated quite simply: the sum of the potential differences across all the components in a closed loop equals zero. To illustrate and clarify this relation, one can consider a single circuit with two sources of electromotive forces E_1 and E_2, and two resistances R_1 and R_2, as shown in Figure 21. The direction chosen for the current i also is indicated. The letters a, b, c, and d are used to indicate certain locations around the circuit. Applying Kirchhoff's second law to the circuit,

$$(V_b - V_a) + (V_c - V_b) + (V_d - V_c) + (V_a - V_d) = 0. \qquad (26)$$

Referring to the circuit in Figure 21, the potential differences maintained by the electromotive forces indicated are $V_b - V_a = E_1$, and $V_c - V_d = -E_2$. From Ohm's law, $V_b - V_c = iR_1$, and $V_d - V_a = iR_2$. Using these four relationships in equation (26), the so-called loop equation becomes $E_1 - E_2 - iR_1 - iR_2 = 0$.

By courtesy of the Department of Physics and Astronomy, Michigan State University

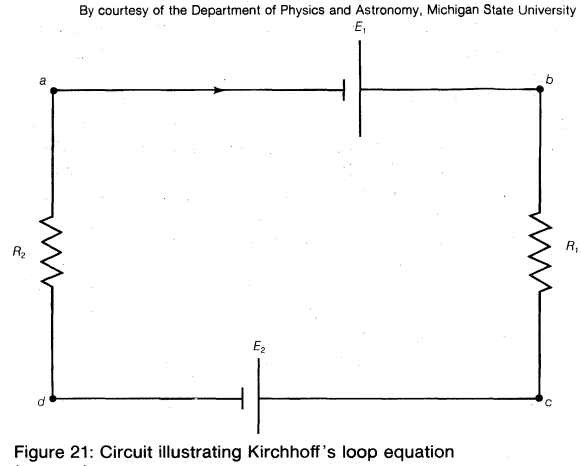


Figure 21: Circuit illustrating Kirchhoff's loop equation (see text).

Given the values of the resistances R_1 and R_2 in ohms and of the electromotive forces E_1 and E_2 in volts, the value of the current i in the circuit is obtained. If E_2 in the circuit had a greater value than E_1, the solution for the current i would be a negative value for i. This negative sign indicates that the current in the circuit would flow in a direction opposite the one indicated in Figure 21.

Kirchhoff's laws can be applied to circuits with several connected loops. The same rules apply, though the algebra required becomes rather tedious as the circuits increase in complexity.

ALTERNATING ELECTRIC CURRENTS

Use in power transmission and in radio and television broadcasting

Basic phenomena and principles. Many applications of electricity and magnetism involve voltages that vary in time. Electric power transmitted over large distances from generating plants to users involves voltages that vary sinusoidally in time, at a frequency of 60 hertz (Hz) in the United States and Canada and 50 hertz in Europe. (One hertz equals one cycle per second.) This means that in the United States, for example, the current alternates its direction in the electric conducting wires so that each second it flows 60 times in one direction and 60 times in the opposite direction. Alternating currents (AC) are also used in radio and television transmissions. In an AM (amplitude-modulation) radio broadcast, electromagnetic waves with a frequency of around one million hertz are generated by currents of the same frequency flowing back and forth in the antenna of the station. The information transported by these waves is encoded in the rapid variation of the wave amplitude. When voices and music are broadcast, these variations correspond to the mechanical oscillations of the sound and have frequencies from 50 to 5,000 hertz. In an FM (frequency-modulation) system, which is used by both television and FM radio stations, audio information is contained in the rapid fluctuation of the frequency in a narrow range around the frequency of the carrier wave.

Circuits that can generate such oscillating currents are called oscillators; they include, in addition to transistors and vacuum tubes, such basic electrical components as resistors, capacitors, and inductors. As was mentioned above, resistors dissipate heat while carrying a current. Capacitors store energy in the form of an electric field in the volume between oppositely charged electrodes. Inductors are essentially coils of conducting wire; they store magnetic energy in the form of a magnetic field generated by the current in the coil. All three components provide some impedance to the flow of alternating currents. In the case of capacitors and inductors, the impedance depends on the frequency of the current. With resistors, impedance is independent of frequency and is simply the resistance. This is easily seen from Ohm's law, equation (21), when it is written as $i = V/R$. For a given voltage difference V between the ends of a resistor, the current varies inversely with the value of R. The greater the value R, the greater is the impedance to the flow of electric current. Before proceeding to circuits with resistors, capacitors, inductors, and sinusoidally varying electromotive forces, the behaviour of a circuit with a resistor and a capacitor will be discussed to clarify transient behaviour and the impedance properties of the capacitor.

Impedance

Transient response. Consider a circuit consisting of a capacitor and a resistor that are connected as shown in Figure 22. What will be the voltage at point b if the voltage at a is increased suddenly from $V_a = 0$ to $V_a = +50$ volts? Closing the switch produces such a voltage because it connects the positive terminal of a 50-volt battery to point a while the negative terminal is at ground (point c). Figure 23 (left) graphs this voltage V_a as a function of the time.

Initially, the capacitor has no charge and does not affect the flow of charge. The initial current is obtained from Ohm's law, $V = iR$, where $V = V_a - V_b$, V_a is 50 volts and V_b is zero. Using 2,000 ohms for the value of the resistance in Figure 22, there is an initial current of 25 milliamperes in the circuit. This current begins to charge the capacitor, so that a positive charge accumulates on the plate of the capacitor connected to point b and a negative charge accumulates on the other plate. As a result, the potential at point b increases from zero to a positive value. As more charge accumulates on the capacitor, this positive potential continues to increase. As it does so, the value of the potential across the resistor is reduced; consequently, the current decreases with time, approaching the value of zero as the capacitor potential reaches 50 volts. The behaviour of the potential at b in Figure 23 (right) is described by the equation $V_b = V_a(1 - e^{-t/RC})$ in volts. For $R = 2{,}000\,\Omega$ and capacitance $C = 2.5$ microfarads, $V_b = 50(1 - e^{-t/0.005})$ in volts. The potential V_b at b in Figure 23 (right) increases from zero when the capacitor is uncharged and reaches the ultimate value of V_a when equilibrium is reached.

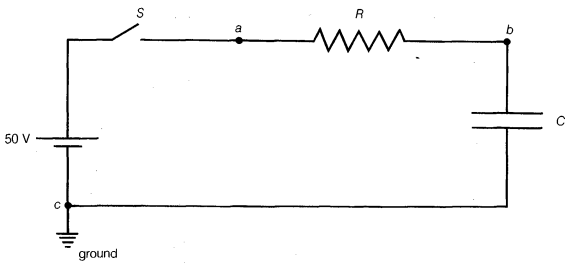


Figure 22: *An RC circuit.*
This type of electric circuit consists of both a resistor and a capacitor connected as shown (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

By courtesy of the Department of Physics and Astronomy, Michigan State University

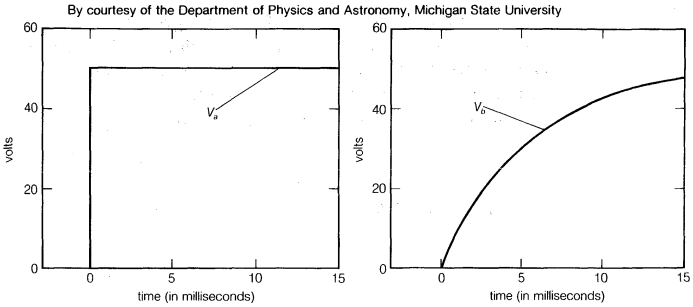


Figure 23: Voltage as a function of time (see text).

How would the potential at point b vary if the potential at point a, instead of being maintained at +50 volts, were to remain at +50 volts for only a short time, say, one millisecond, and then return to zero? The superposition principle (see above) is used to solve the problem. The voltage at a starts at zero, goes to +50 volts at $t = 0$, then returns to zero at $t = +0.001$ second. This voltage can be viewed as the sum of two voltages, $V_{1a} + V_{2a}$, where V_{1a} becomes +50 volts at $t = 0$ and remains there indefinitely, and V_{2a} becomes −50 volts at $t = 0.001$ second and remains there indefinitely. This superposition is shown graphically on the left side of Figure 24. Since the solutions for V_{1b} and V_{2b} corresponding to V_{1a} and V_{2a} are known from the previous example, their sum V_b is the answer to the problem. The individual solutions and their sum are given graphically on the right side of Figure 24.

The voltage at b reaches a maximum of only 9 volts. The superposition illustrated in Figure 24 also shows that the shorter the duration of the positive "pulse" at a, the smaller is the value of the voltage generated at b. Increasing the size of the capacitor also decreases the maximum voltage at b. This decrease in the potential of a transient explains the "guardian role" that capacitors play in protecting delicate and complex electronic circuits from damage by large transient voltages. These transients, which generally occur at high frequency, produce effects similar to those produced by pulses of short duration. They can damage equipment when they induce circuit components to break down electrically. Transient voltages are often introduced into electronic circuits through power supplies. A concise way to describe the role of the capacitor in the above example is to say that its impedance to an electric signal decreases with increasing frequency. In the example, much of the signal is shunted to ground instead of appearing at point b.

Transient voltages and their effects

Alternating-current circuits. Certain circuits include

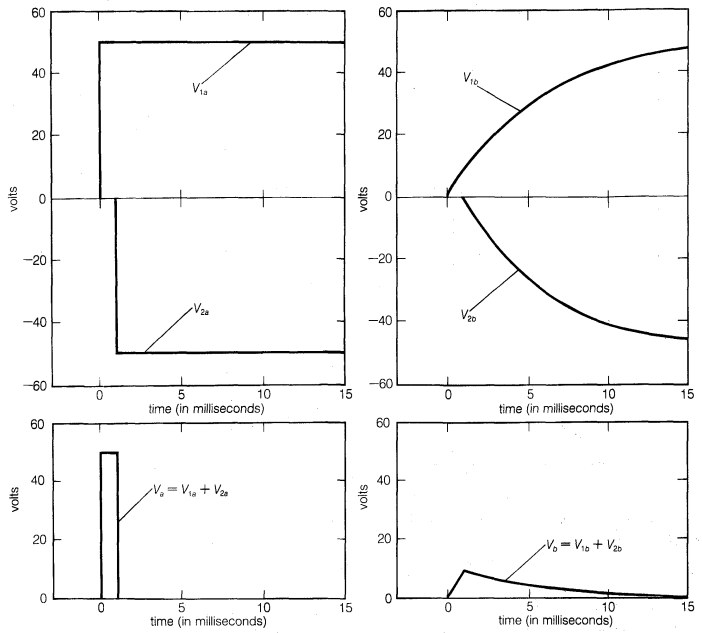


Figure 24: Application of the superposition principle to a problem concerned with voltages as a function of time (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

sources of alternating electromotive forces of the sinusoidal form $V = V_0 \cos(\omega t)$ or $V = V_0 \sin(\omega t)$. The sine and cosine functions have values that vary between +1 and −1; either of the equations for the voltage represents a potential that varies with respect to time and has values from $+V_0$ to $-V_0$. The voltage varies with time at a rate given by the numerical value of ω; ω, which is called the angular frequency, is expressed in radians per second. Figure 25 shows an example with $V_0 = 170$ volts and $\omega = 377$ radians per second, so that $V = 170 \cos(377t)$. The time interval required for the pattern to be repeated is called the period T, given by $T = 2\pi/\omega$. In Figure 25, the pattern is repeated every 16.7 milliseconds, which is the period. The frequency of the voltage is symbolized by f and given by $f = 1/T$. In terms of ω, $f = \omega/2\pi$, in hertz.

The root-mean-square (rms) voltage of a sinusoidal source of electromotive force (V_{rms}) is used to characterize the source. It is the square root of the time average of the voltage squared. The value of V_{rms} is $V_0/\sqrt{2}$, or, equivalently, $0.707V_0$. Thus, the 60-hertz, 120-volt alternating current, which is available from most electric outlets in U.S. homes and which is illustrated in Figure 25, has $V_0 = 120/0.707 = 170$ volts. The potential difference at the outlet varies from +170 volts to −170 volts and back to +170 volts 60 times each second. The rms values of voltage and current are especially useful in calculating average power in AC circuits.

A sinusoidal electromotive force can be generated using

By courtesy of the Department of Physics and Astronomy, Michigan State University

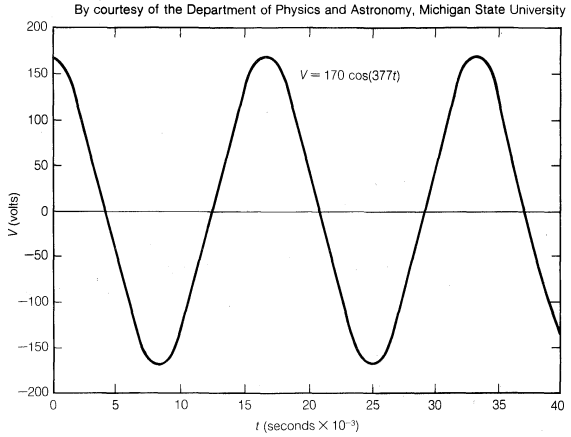


Figure 25: A sinusoidal voltage (see text).

the principles described in Faraday's law of electromagnetic induction (see below *Faraday's law of induction*). Briefly, an alternating electromotive force can be induced in a loop of conducting wire by rotating the loop of wire in a uniform magnetic field.

In AC circuits, it is often necessary to find the currents as a function of time in the various parts of the circuit for a given source of sinusoidal electromotive force. While the problems can become quite complex, the solutions are based on Kirchhoff's two laws discussed above (see *Kirchhoff's laws of electric circuits*). The solution for the current in a given loop takes the form $i = i_0 \cos(\omega t - \varphi)$. The current has the same frequency as the applied voltage but is not necessarily "in phase" with that voltage. When the phase angle φ does not equal zero, the maximum of the current does not occur when the driving voltage is at its maximum.

Behaviour of an AC circuit

The way an AC circuit functions can be better understood by examining one that includes a source of sinusoidally varying electromotive force, a resistor, a capacitor, and an inductor, all connected in series. For this single-loop problem, only the second of Kirchhoff's laws is needed since there is only one current. The circuit is shown in Figure 26 with the points *a, b, c,* and *d* at various positions in the circuit located between the various elements. The letters *R, L,* and *C* represent, respectively, the values of

By courtesy of the Department of Physics and Astronomy, Michigan State University

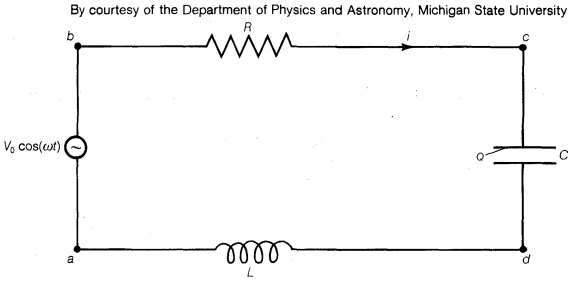


Figure 26: *A series LRC circuit.*
This type of electric circuit has an inductor, resistor, and capacitor connected in series (see text).

the resistance in ohms, the inductance in henrys, and the capacitance in farads. The source of the AC electromotive force is located between *a* and *b*. The wavy symbol is a reminder of the sinusoidal nature of the voltage that is responsible for making the current flow in the loop. For the potential between *b* and *a*,

$$V_b - V_a = V_0 \cos \omega t. \qquad (27a)$$

Equation (27a) represents a potential difference that has its maximum positive value at $t = 0$.

The direction chosen for the current i in the circuit in Figure 26 represents the direction of that current at some particular time, since AC circuits feature continuous reversals of the direction of the flow of charge. The direction chosen for the current is important, however, because the loop equation must consider all the elements at the same instant in time. The potential difference across the resistor is given by Ohm's law as

$$V_b - V_c = iR. \qquad (27b)$$

For equation (27b), the direction of the current is important. The potential difference across the capacitor, $V_c - V_d$, depends on the charge on the capacitor. When the charge on the upper plate of the capacitor in Figure 26 has a value Q, the potential difference across the capacitor is

$$V_c - V_d = \frac{Q}{C}, \qquad (27c)$$

which is a variant of equation (12). One must be careful labeling the charge and the direction of the current, since the charge on the other plate is $-Q$. For the choices shown in the figure, the current in the circuit is given by the rate of change of the charge Q—that is, $i = dQ/dt$. Finally, the value of the potential difference $V_d - V_a$ across the inductor depends on the rate of change of the current through the inductor, di/dt. For the direction chosen for i, the value is

$$V_d - V_a = +L\frac{di}{dt}. \tag{27d}$$

The result of combining equations (27a, b, c, d) in accordance with Kirchhoff's second law for the loop in Figure 26 is

$$V_0 \cos(\omega t) = L\frac{di}{dt} + iR + \frac{Q}{C}. \tag{28}$$

Both the current i and the rate of change of the current di/dt can be eliminated from equation (28), since $i = dQ/dt$, and $di/dt = d^2Q/dt^2$. The result is a linear, inhomogeneous, second-order differential equation with well-known solutions for the charge Q as a function of time. The most important solution describes the current and voltages after transient effects have been dampened; the transient effects last only a short time after the circuit is completed. Once the charge is known, the current in the circuit can be obtained by taking the first derivative of the charge. The expression for the current in the circuit is

$$i = \frac{V_0}{Z}\cos(\omega t - \varphi) = i_0 \cos(\omega t - \varphi). \tag{29}$$

In equation (29), Z is the impedance of the circuit; impedance, like resistance, is measured in units of ohms. Z is a function of the frequency of the source of applied electromotive force. The equation for Z is

$$Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}. \tag{30}$$

If the resistor were the only element in the circuit, the impedance would be $Z = R$, the resistance of the resistor. For a capacitor alone, $Z = 1/\omega C$, showing that the impedance of a capacitor decreases as the frequency increases. For an inductor alone, $Z = \omega L$; the reason why the impedance of the inductor increases with frequency will become clear once Faraday's law of magnetic induction is discussed in detail below. Here it is sufficient to say that an induced electromotive force in the inductor opposes the change in current, and it is directly proportional to the frequency.

The phase angle φ in equation (29) gives the time relationship between the current in the circuit and the driving electromotive force, $V_0 \cos(\omega t)$. The tangent of the angle φ is

$$\tan\varphi = \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R}. \tag{31}$$

Depending on the values of ω, L, and C, the angle φ can be positive, negative, or zero. If φ is positive, the current "lags" the voltage, while for negative values of φ, the current "leads" the voltage.

The power dissipated in the circuit is the same as the power delivered by the source of electromotive force, and both are measured in watts. Using equation (23), the power is given by

$$P = iV = i_0 \cos(\omega t - \varphi) V_0 \cos(\omega t). \tag{32}$$

An expression for the average power dissipated in the circuit can be written either in terms of the peak values i_0 and V_0 or in terms of the rms values i_{rms} and V_{rms}. The average power is

$$P_{ave} = I_{rms} V_{rms} \cos\varphi = \frac{1}{2} i_0 V_0 \cos\varphi. \tag{33}$$

The cos φ in equation (33) is called the power factor. It is evident that the only element that can dissipate energy is the resistance.

Resonance

A most interesting condition known as resonance occurs when the phase angle is zero in equation (31), or equivalently, when the angular frequency ω has the value $\omega = \omega_r = \sqrt{1/LC}$. The impedance in equation (30) then has its minimum value and equals the resistance R. The amplitude of the current in the circuit, i_0, is at its maximum value (see equation [29]). Figure 27 shows the dependence of i_0 on the angular frequency ω of the source of alternating electromotive force. The values of the electric parameters for the figure are $V_0 = 50$ volts, $R = 25$ ohms,

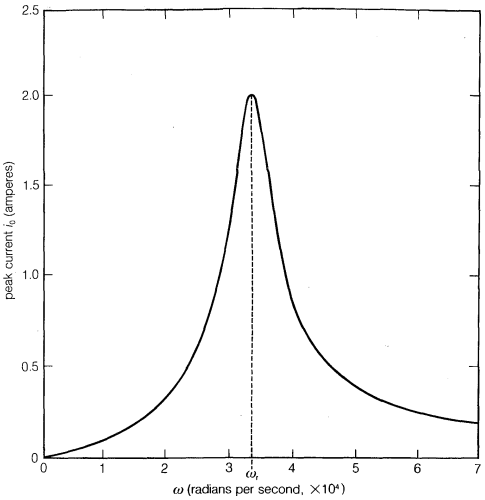


Figure 27: Current amplitude (peak current) as a function of ω (see text).

By courtesy of the Department of Physics and Astronomy, Michigan State University

$L = 4.5$ millihenrys, and $C = 0.2$ microfarad. With these values, the resonant angular frequency ω_r of the circuit in Figure 26 is 3.33×10^4 radians per second.

The peaking in the current shown in Figure 27 constitutes a resonance. At the resonant frequency, in this case when ω_r equals 3.33×10^4 radians per second, the impedance Z of the circuit is at a minimum and the power dissipated is at a maximum. The phase angle φ is zero so that the current is in phase with the driving voltage, and the power factor, cos φ, is 1. Figure 28 illustrates the variation of the average power with the angular frequency of the sinusoidal electromotive force. The resonance is seen to be even more pronounced. The quality factor Q for the circuit is the electric energy stored in the circuit divided by the energy dissipated in one period. The Q of a circuit is an important quantity in certain applications, as in the case of electromagnetic waveguides and radio-frequency cavities where Q has values around 10,000 and where high voltages and electric fields are desired. For the present circuit, $Q = \omega_r L/R$. Q also can be obtained from the average power graph as the ratio $\omega_r/(\omega_2 - \omega_1)$, where ω_1 and ω_2 are the angular frequencies at which the average power dissipated in the circuit is one-half its maximum value. For the circuit here, $Q = 6$.

By courtesy of the Department of Physics and Astronomy, Michigan State University

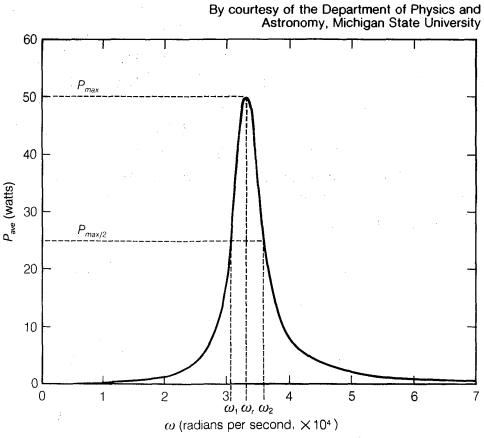


Figure 28: Average power dissipation versus ω (see text).

What is the maximum value of the potential difference across the inductor? Since it is given by Ldi/dt, it will occur when the current has the maximum rate of change. Figure 29 shows the amplitude of the potential difference as a function of ω.

The maximum amplitude of the voltage across the inductor, 300 volts, is much greater than the 50-volt amplitude of the driving sinusoidal electromotive force. This result is typical of resonance phenomena. In a familiar mechanical system, children on swings time their kicks to attain very

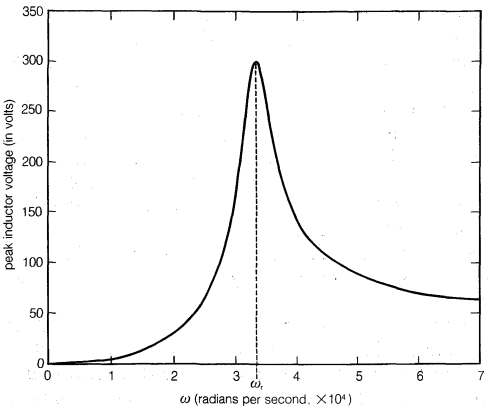


Figure 29: Electromotive force across L versus ω (see text).

By courtesy of the Department of Physics and Astronomy, Michigan State University

large swings (much larger than they could attain with a single kick). In a more spectacular, albeit costly, example, the collapse of the Tacoma Narrows Bridge (a suspension bridge across the Narrows of Puget Sound, Wash.) on Nov. 7, 1940, was the result of the large amplitudes of oscillations that the span attained as it was driven in resonance by high winds. A ubiquitous example of electric resonance occurs when a radio dial is turned to receive a broadcast. Turning the dial changes the value of the tuning capacitor of the radio. When the circuit attains a resonance frequency corresponding to the frequency of the radio wave, the voltage induced is enhanced and processed to produce sound.

Magnetism

FUNDAMENTALS

Magnetism is a phenomenon associated with the motion of charge. This motion can take many forms. It can be an electric current in a conductor or charged particles moving through space, or it can be the motion of an electron in atomic orbit. Magnetism is also associated with elementary particles, such as the electron, that have a property called spin.

Effects of a magnetic field

Basic to magnetism are magnetic fields and their effects on matter, as, for instance, the deflection of moving charges and torques on other magnetic objects. Evidence for the presence of a magnetic field is the magnetic force on charges moving in that field; the force is at right angles to both the field and the velocity of the charge. This force deflects the particles without changing their speed. The deflection can be observed in the electron beam of a television tube when a permanent magnet is brought near the tube. A more familiar example is the torque on a compass needle that acts to align the needle with the magnetic field of the Earth. The needle is a thin piece of iron that has been magnetized—*i.e.*, a small bar magnet. One end of the magnet is called a north pole and the other end a south pole. The force between a north and a south pole is attractive, whereas the force between like poles is repulsive. The magnetic field is sometimes referred to as magnetic induction or magnetic flux density; it is always symbolized by $\boldsymbol{B}$. Magnetic fields are measured in units of tesla (T). (Another unit of measure commonly used for $\boldsymbol{B}$ is the gauss, though it is no longer considered a standard unit. One gauss equals 10^{-4} tesla.)

A fundamental property of a magnetic field is that its flux through any closed surface vanishes. (A closed surface is one that completely surrounds a volume.) This is expressed mathematically by div $\boldsymbol{B} = 0$ and can be understood physically in terms of the field lines representing $\boldsymbol{B}$. These lines always close on themselves, so that if they enter a certain volume at some point, they must also leave that volume. In this respect, a magnetic field is quite different from an electric field. Electric field lines can begin and end on a charge, but no equivalent magnetic charge has been found in spite of many searches for so-called magnetic monopoles.

The most common source of magnetic fields is the electric current loop. It may be an electric current in a circular conductor or the motion of an orbiting electron in an atom. Associated with both these types of current loops is a magnetic dipole moment, the value of which is iA, the product of the current and the area of the loop. In addition, electrons, protons, and neutrons in atoms have a magnetic dipole moment associated with their intrinsic spin; such magnetic dipole moments represent another important source of magnetic fields. A particle with a magnetic dipole moment is often referred to as a magnetic dipole. (A magnetic dipole may be thought of as a tiny bar magnet. It has the same magnetic field as such a magnet and behaves the same way in external magnetic fields.) When placed in an external magnetic field, a magnetic dipole can be subjected to a torque that tends to align it with the field; if the external field is not uniform, the dipole also can be subjected to a force.

All matter exhibits magnetic properties to some degree. When placed in an inhomogeneous field, matter is either attracted or repelled in the direction of the gradient of the field. This property is described by the magnetic susceptibility of the matter and depends on the degree of magnetization of the matter in the field. Magnetization depends on the size of the dipole moments of the atoms in a substance and the degree to which the dipole moments are aligned with respect to each other. Certain materials, such as iron, exhibit very strong magnetic properties because of the alignment of the magnetic moments of their atoms within certain small regions called domains. Under normal conditions, the various domains have fields that cancel, but they can be aligned with each other to produce extremely large magnetic fields. Various alloys, like NdFeB (an alloy of neodymium, iron, and boron), keep their domains aligned and are used to make permanent magnets. The strong magnetic field produced by a typical three-millimetre-thick magnet of this material is comparable to an electromagnet made of a copper loop carrying a current of several thousand amperes. In comparison, the current in a typical light bulb is 0.5 ampere. Since aligning the domains of a material produces a magnet, disorganizing the orderly alignment destroys the magnetic properties of the material. Thermal agitation that results from heating a magnet to a high temperature destroys its magnetic properties.

Domains

Magnetic fields vary widely in strength. Some representative values are given in Table 3.

Table 3: Typical Magnetic Fields

Inside atomic nuclei	10^{11} T
In superconducting solenoids	20 T
In a superconducting coil cyclotron	5 T
Near a small ceramic magnet	0.1 T
Earth's field at the equator	4×10^{-5} T
In interstellar space	2×10^{-10} T

MAGNETIC FIELD OF STEADY CURRENTS

Biot–Savart law

Magnetic fields produced by electric currents can be calculated for any shape of circuit using the law of Biot and Savart, named for the early 19th-century French physicists Jean-Baptiste Biot and Félix Savart. A few magnetic field lines produced by a current in a loop are shown in Figure 30. These lines of $\boldsymbol{B}$ form loops around the current. The Biot–Savart law expresses the partial contribution $d\boldsymbol{B}$ from a small segment of conductor to the total $\boldsymbol{B}$ field of a current in the conductor. For a segment of length and orientation $d\boldsymbol{l}$ that carries a current i,

$$d\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{i d\boldsymbol{l} \times \hat{\boldsymbol{r}}}{r^2}. \qquad (34)$$

In this equation, μ_0 is the permeability of free space and has the value of $4\pi \times 10^{-7}$ newton per square ampere. This equation is illustrated in Figure 31 for a small segment of a wire that carries a current so that, at the origin of the coordinate system, the small segment of length $d\boldsymbol{l}$ of the wire lies along the x axis.

Comparing $d\boldsymbol{B}$ at points 1 and 2 shows the inverse square dependence of the magnitude of the field with distance. The vectors at points 1, 3, and 4, which are all at the

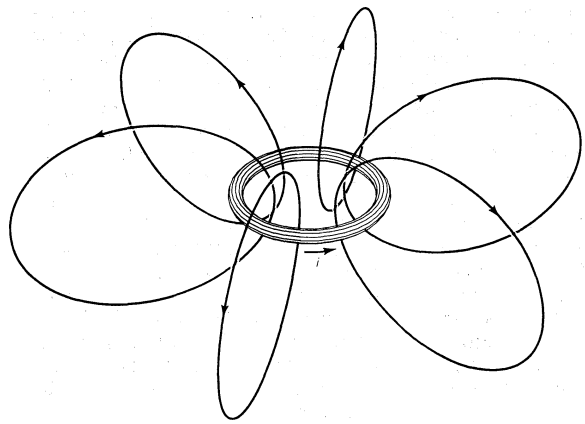


Figure 30: Some lines of the magnetic field $\boldsymbol{B}$ for an electric current i in a loop (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

same distance from $d\boldsymbol{l}$, show the direction of $d\boldsymbol{B}$ in a circle around the wire. In position 1, the contribution to the field, $d\boldsymbol{B}_1$, is perpendicular both to the current direction and to the vector r_1. Finally, the vectors at 1, 5, 6, and 7 illustrate the angular dependence of the magnitude of $d\boldsymbol{B}$ at a point. The magnitude of $d\boldsymbol{B}$ varies as the sine of the angle between $d\boldsymbol{l}$ and $\hat{\boldsymbol{r}}$, where $\hat{\boldsymbol{r}}$ is in the direction from $d\boldsymbol{l}$ to the point. It is strongest at 90° to $d\boldsymbol{l}$ and decreases to zero for locations directly in line with $d\boldsymbol{l}$. The magnetic field of a current in a loop or coil is obtained by summing the individual partial contributions of all the segments of the circuits, taking into account the vector nature of the field. While simple mathematical expressions for the magnetic field can be derived for a few current configurations, most of the practical applications require the use of high-speed computers.

The expression for the magnetic field $\boldsymbol{B}$ a distance r from a long straight wire with current i is

$$\boldsymbol{B} = \frac{\mu_0 i}{2\pi r}\hat{\boldsymbol{\theta}}, \tag{35}$$

where $\hat{\boldsymbol{\theta}}$ is a unit vector pointing in a circle around the wire. The $\boldsymbol{B}$ field near a long straight wire with current i can be seen in Figures 2A and 2B. The magnetic field at a distance r from a magnetic dipole with moment $\boldsymbol{m}$ is given by

$$\boldsymbol{B} = \frac{\mu_0 m}{4\pi r^3}(2\cos\theta\,\hat{\boldsymbol{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}). \tag{36}$$

The size of the magnetic dipole moment is m in ampere times square metre (A · m²), and the angle between the direction of $\boldsymbol{m}$ and of $\boldsymbol{r}$ is θ. Both $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{\theta}}$ are unit vectors in the direction of $\boldsymbol{r}$ and θ. It is apparent that the magnetic

By courtesy of the Department of Physics and Astronomy, Michigan State University

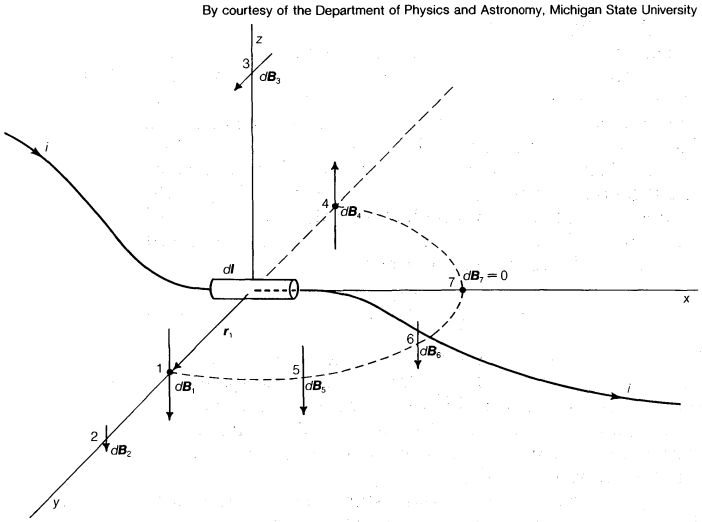


Figure 31: A magnetic field produced by a small section of wire with electric current i (see text).

field decreases rapidly as the cube of the distance from the dipole. Equation (36) is also valid for a small current loop with current i, when the distance r is much greater than the size of the current loop. A loop of area A has a magnetic dipole moment with a magnitude $m = iA$; its direction is perpendicular to the plane of the loop, along the direction of $\boldsymbol{B}$ inside the loop. If the fingers of the right hand are curled and held in the direction of the current in the loop, the extended thumb points in the direction of $\boldsymbol{m}$. In Figure 30, the dipole moment of the current in the loop points up; in Figure 32, $\boldsymbol{m}$ points down because the current flows in a clockwise direction when viewed from above.

The magnetic field of the current loop in Figure 32 at points far from the loop has the same shape as the electric field of an electric dipole; the latter consists of two equal charges of opposite sign separated by a small distance. Magnetic dipoles, like electric dipoles, occur in a variety of situations. Electrons in atoms have a magnetic dipole moment that corresponds to the current of their orbital motion around the nucleus. In addition, the electrons have a magnetic dipole moment associated with their spin. The Earth's magnetic field is thought to be the result of cur-

By courtesy of the Department of Physics and Astronomy, Michigan State University

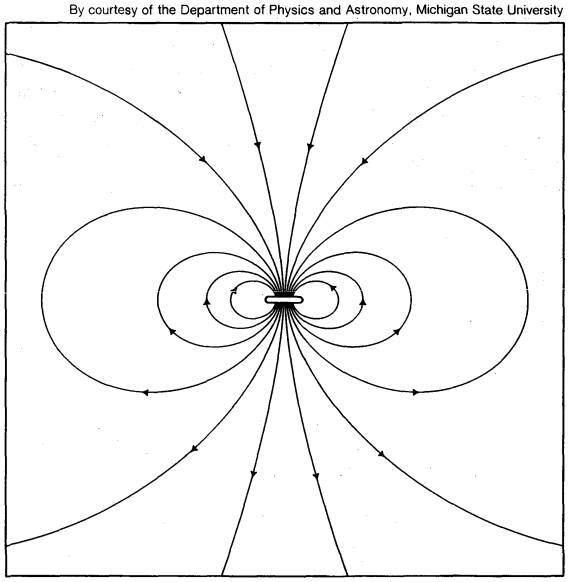


Figure 32: Some of the lines of $\boldsymbol{B}$ from the small current loop in the centre. The current in the loop flows in a clockwise direction when viewed from above.

rents related to the planet's rotation. The magnetic field far from a small bar magnet is well represented by the field of a magnetic dipole. In most of these cases, moving charge produces a magnetic field $\boldsymbol{B}$. Inside a long solenoid with current i and away from its ends, the magnetic field is uniform and directed along the axis of the solenoid. A solenoid of this kind can be made by wrapping some conducting wire tightly around a long hollow cylinder. The value of the field is

$$\boldsymbol{B} = \mu_0 n i, \tag{37}$$

where n is the number of turns per unit length of the solenoid.

MAGNETIC FORCES

A magnetic field $\boldsymbol{B}$ imparts a force on moving charged particles. The entire electromagnetic force on a charged particle with charge q and velocity $\boldsymbol{v}$ is called the Lorentz force (after the Dutch physicist Hendrik A. Lorentz) and is given by

Lorentz force

$$\boldsymbol{F} = q\boldsymbol{E} + q\boldsymbol{v} \times \boldsymbol{B}. \tag{38}$$

The first term is contributed by the electric field. The second term is the magnetic force and has a direction perpendicular to both the velocity $\boldsymbol{v}$ and the magnetic field $\boldsymbol{B}$. The magnetic force is proportional to q and to the magnitude of $\boldsymbol{v} \times \boldsymbol{B}$. In terms of the angle φ between $\boldsymbol{v}$ and $\boldsymbol{B}$, the magnitude of the force equals $qvB \sin \varphi$. An interest-

ing result of the Lorentz force is the motion of a charged particle in a uniform magnetic field. If v is perpendicular to B (*i.e.*, with the angle φ between v and B of 90°), the particle will follow a circular trajectory with a radius of $r = mv/qB$. If the angle φ is less than 90°, the particle orbit will be a helix with an axis parallel to the field lines. If φ is zero, there will be no magnetic force on the particle, which will continue to move undeflected along the field lines. Charged particle accelerators like cyclotrons make use of the fact that particles move in a circular orbit when v and B are at right angles. For each revolution, a carefully timed electric field gives the particles additional kinetic energy, which makes them travel in increasingly larger orbits. When the particles have acquired the desired energy, they are extracted and used in a number of different ways, from fundamental studies of the properties of matter to the medical treatment of cancer.

The magnetic force on a moving charge reveals the sign of the charge carriers in a conductor. A current flowing from right to left in a conductor can be the result of positive charge carriers moving from right to left or negative charges moving from left to right, or some combination of each. When a conductor is placed in a B field perpendicular to the current, the magnetic force on both types of charge carriers is in the same direction. This force, which can be seen in Figure 3, gives rise to a small potential difference between the sides of the conductor. Known as the

Hall effect

Hall effect, this phenomenon (discovered by the American physicist Edwin H. Hall) results when an electric field is aligned with the direction of the magnetic force. As is evident in Figure 3, the sign of the potential differs according to the sign of the charge carrier because, in one case, positive charges are pushed toward the reader and, in the other, negative charges are pushed in that direction. The Hall effect shows that electrons dominate the conduction of electricity in copper. In zinc, however, conduction is dominated by the motion of positive charge carriers. Electrons in zinc that are excited from the valence band leave holes, which are vacancies (*i.e.*, unfilled levels) that behave like positive charge carriers. The motion of these holes accounts for most of the conduction of electricity in zinc.

If a wire with a current i is placed in an external magnetic field B, how will the force on the wire depend on the orientation of the wire? Since a current represents a movement of charges in the wire, the Lorentz force given in equation (38) acts on the moving charges. Because these charges are bound to the conductor, the magnetic forces on the moving charges are transferred to the wire. The force on a small length dl of the wire depends on the orientation of the wire with respect to the field. The magnitude of the force is given by $idlB \sin \varphi$, where φ is the angle between B and dl. There is no force when $\varphi = 0$ or 180°, both of which correspond to a current along a direction parallel to the field. The force is at a maximum when the current and field are perpendicular to each other. The force is obtained from equation (38) and is given by

$$dF = idl \times B. \qquad (39)$$

Again, the cross product denotes a direction perpendicular to both dl and B. The direction of dF is given by the right-hand rule illustrated in Figure 33. As shown, the fingers are in the direction of B; the current (or in the case of a positive moving point charge, the velocity) is in the direction of the thumb, and the force is perpendicular to the palm.

By courtesy of the Department of Physics and Astronomy, Michigan State University

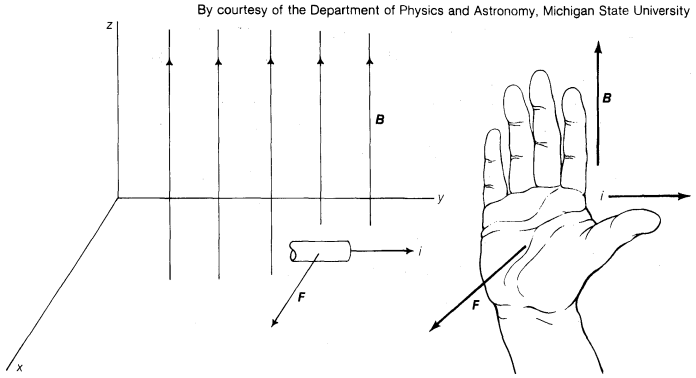


Figure 33: Right-hand rules for the magnetic force on an electric current (see text).

The force between two wires, each of which carries a current, can be understood from the interaction of one of the currents with the magnetic field produced by the other current. For example, the force between two parallel wires carrying currents in the same direction is attractive. It is repulsive if the currents are in opposite directions. Two circular current loops, located one above the other and with their planes parallel, will attract if the currents are in the same directions and will repel if the currents are in opposite directions. The situation is shown on the left side of Figure 34. When the loops are side by side as on the right side of Figure 34, the situation is reversed. For two currents flowing in the same direction, whether clockwise or counterclockwise, the force is repulsive, while for opposite directions, it is attractive. The nature of the force for the loops depicted in Figure 34 can be obtained by considering the direction of the currents in the parts of the loops that are closest to each other: same current direction, attraction; opposite current direction, repulsion. This seemingly complicated force between current loops can be understood more simply by treating the fields as though they originated from magnetic dipoles. As discussed above, the B field of a small current loop is well represented by the field of a magnetic dipole at distances that are large compared to the size of the loop. In another way of looking at the interaction of current loops, the loops of Figure 34A and 34B are replaced in Figure 35A and 35B by small permanent magnets, with the direction of the magnets from south to north corresponding to the direction of the magnetic moment of the loop m. Outside the magnets, the magnetic field lines point away from the north pole and toward the south pole.

By courtesy of the Department of Physics and Astronomy, Michigan State University

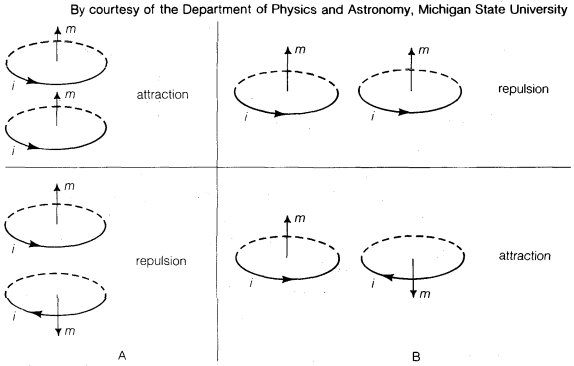


Figure 34: *Magnetic force between current loops.* In each case shown, the arrow indicates the direction of the current i and the magnetic dipole moment m of a loop (see text).

It is easy to understand the nature of the forces in Figures 34 and 35 with the rule that two north poles repulse each other and two south poles repulse each other, while unlike poles attract. As was noted earlier, Coulomb established an inverse square law of force for magnetic poles and electric charges; according to his law, unlike poles attract and like poles repel, just as unlike charges attract and like charges repel. Today, Coulomb's law refers only to charges, but historically it provided the foundation for a magnetic potential analogous to the electric potential.

The alignment of a magnetic compass needle with the direction of an external magnetic field is a good example of the torque to which a magnetic dipole is subjected. The torque has a magnitude $\tau = mB \sin \vartheta$. Here, ϑ is the angle between m and B. The torque τ tends to align m with B. It has its maximum value when ϑ is 90°, and it is zero when the dipole is in line with the external field. Rotating a magnetic dipole from a position where $\vartheta = 0$ to a position where $\vartheta = 180°$ requires work. Thus, the potential energy of the dipole depends on its orientation with respect to the field and is given in units of joules by

$$U_m = -mB \cos \vartheta. \qquad (40)$$

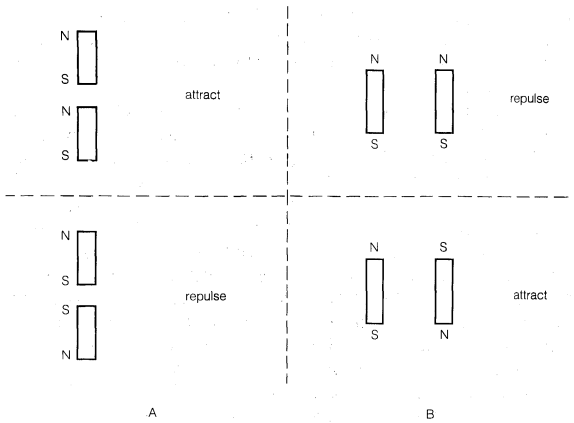


Figure 35: Force between small permanent bar magnets.
By courtesy of the Department of Physics and Astronomy, Michigan State University

Basis for magnetic resonance imaging

Equation (40) represents the basis for an important medical application—namely, magnetic resonance imaging (MRI), also known as nuclear magnetic resonance imaging. MRI involves measuring the concentration of certain atoms, most commonly those of hydrogen, in body tissue and processing this measurement data to produce high-resolution images of organs and other anatomical structures. When hydrogen atoms are placed in a magnetic field, their nuclei (protons) tend to have their magnetic moments preferentially aligned in the direction of the field. The magnetic potential energy of the nuclei is calculated according to equation (40) as $-mB$. Inverting the direction of the dipole moment requires an energy of $2mB$, since the potential energy in the new orientation is $+mB$. A high-frequency oscillator provides energy in the form of electromagnetic radiation of frequency v, with each quantum of radiation having an energy hv, where h is Planck's constant. The electromagnetic radiation from the oscillator consists of high-frequency radio waves, which are beamed into the patient's body while it is subjected to a strong magnetic field. When the resonance condition $hv = 2mB$ is satisfied, the hydrogen nuclei in the body tissue absorb the energy and reverse their orientation. The resonance condition is met in only a small region of the body at any given time, and measurement of the energy absorption reveals the concentration of hydrogen atoms in that region alone. The magnetic field in an MRI scanner is usually provided by a large solenoid with B of one to three teslas. A number of "gradient coils" insures that the resonance condition is satisfied solely in the limited region inside the solenoid at any particular time; the coils are used to move this small target region, thereby making it possible to scan the patient's body throughout. The frequency of the radiation v is determined by the value of B and is typically 40 to 130 megahertz. The MRI technique does not harm the patient because the energy of the quanta of the electromagnetic radiation is much smaller than the thermal energy of a molecule in the human body.

The direction of the magnetic moment $\boldsymbol{m}$ of a compass needle is from the end marked S for south to the one marked N for north. The lowest energy occurs for $\vartheta = 0$, when $\boldsymbol{m}$ and $\boldsymbol{B}$ are aligned. In a typical situation, the compass needle comes to rest after a few oscillations and points along the $\boldsymbol{B}$ field in the direction called north. It must be concluded from this that the Earth's North Pole is really a magnetic south pole, with the field lines pointing toward that pole, while its South Pole is a magnetic north pole. Put another way, the dipole moment of the Earth currently points north to south. Short-term changes in the Earth's magnetic field are ascribed to electric currents in the ionosphere. There are also longer-term fluctuations in the locations of the poles. The angle between the compass needle and geographic north is called the magnetic declination (see EARTH: *The magnetic field of the Earth*).

The repulsion or attraction between two magnetic dipoles can be viewed as the interaction of one dipole with the magnetic field produced by the other dipole. The magnetic field is not constant, but varies with the distance from the dipole. When a magnetic dipole with moment $\boldsymbol{m}$ is in a $\boldsymbol{B}$ field that varies with position, it is subjected to a force proportional to that variation—*i.e.*, to the gradient of $\boldsymbol{B}$. The direction of the force is understood best by considering the potential energy of a dipole in an external $\boldsymbol{B}$ field, as given by equation (40). The force on the dipole is in the direction in which that energy decreases most rapidly. For example, if the magnetic dipole $\boldsymbol{m}$ is aligned with $\boldsymbol{B}$, then the energy is $-mB$, and the force is in the direction of increasing $\boldsymbol{B}$. If $\boldsymbol{m}$ is directed opposite to $\boldsymbol{B}$, then the potential energy given by equation (40) is $+mB$, and in this case the force is in the direction of decreasing $\boldsymbol{B}$. Both types of forces are observed when various samples of matter are placed in a nonuniform magnetic field. Such a field from an electromagnet is sketched in Figure 36.

Regardless of the direction of the magnetic field in Figure 36, a sample of copper is magnetically attracted toward the low field region to the right in the drawing. This behaviour is termed diamagnetism. A sample of aluminum, however, is attracted toward the high field region in an effect called paramagnetism. A magnetic dipole moment is induced when matter is subjected to an external field.

By courtesy of the Department of Physics and Astronomy, Michigan State University

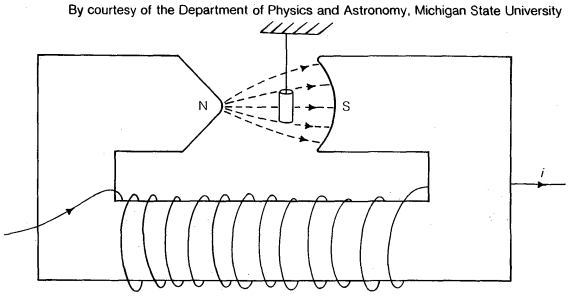


Figure 36: A small sample of copper in an inhomogeneous magnetic field (see text).

For copper, the induced dipole moment is opposite to the direction of the external field; for aluminum, it is aligned with that field. The magnetization $\boldsymbol{M}$ of a small volume of matter is the sum (a vector sum) of the magnetic dipole moments in the small volume divided by that volume. $\boldsymbol{M}$ is measured in units of amperes per metre. The degree of induced magnetization is given by the magnetic susceptibility of the material χ_m, which is commonly defined by the equation

$$\boldsymbol{M} = \chi_m \boldsymbol{H}. \tag{41}$$

The field $\boldsymbol{H}$ is called the magnetic intensity and, like $\boldsymbol{M}$, is measured in units of amperes per metre. (It is sometimes also called the magnetic field, but the symbol $\boldsymbol{H}$ is unambiguous.) The definition of $\boldsymbol{H}$ is

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M}. \tag{42}$$

Magnetization effects in matter are discussed in some detail below. The permeability μ is often used for ferromagnetic materials such as iron that have a large magnetic susceptibility dependent on the field and the previous magnetic state of the sample; permeability is defined by the equation $\boldsymbol{B} = \mu \boldsymbol{H}$. From equations (41) and (42), it follows that $\mu = \mu_0 (1 + \chi_m)$.

The effect of ferromagnetic materials in increasing the magnetic field produced by current loops is quite large. Figure 37 illustrates a toroidal winding of conducting wire around a ring of iron that has a small gap. The magnetic field inside a toroidal winding similar to the one illustrated in Figure 37 but without the iron ring is given by $B = \mu_0 Ni/2\pi r$, where r is the distance from the axis of the toroid, N is the number of turns, and i is the current in the wire. The value of B for $r = 0.1$ metre, $N = 100$, and $i = 10$ amperes is only 0.002 tesla—about 50 times the magnetic field at the Earth's surface. If the same toroid is wound around an iron ring with no gap, the magnetic field inside the iron is larger by a factor equal to μ/μ_0, where μ is the magnetic permeability of the iron. For low-carbon iron in these conditions, $\mu = 8{,}000\mu_0$. The magnetic field in the iron is then 1.6 tesla. In a typical electromagnet, iron is used to increase the field in a small region, such

as the narrow gap in the iron ring illustrated in Figure 37. If the gap is one centimetre wide, the field in that gap is about 0.12 tesla, a 60-fold increase relative to the 0.002-tesla field in the toroid when no iron is used. This factor is typically given by the ratio of the circumference of the toroid to the gap in the ferromagnetic material. The maximum value of B as the gap becomes very small is of course the 1.6 tesla obtained above when there is no gap.

Energy density in a magnetic field

The energy density in a magnetic field is given in the absence of matter by $\frac{1}{2}B^2/\mu_0$; it is measured in units of joules per cubic metre. The total magnetic energy can be obtained by integrating the energy density over all space. The direction of the magnetic force can be deduced in many situations by studying distribution of the magnetic field lines; motion is favoured in the direction that tends to decrease the volume of space where the magnetic field is strong. This can be understood because the magnitude of B is squared in the energy density. Figure 38 shows some lines of the B field for two circular current loops with currents in opposite directions.

By courtesy of the Department of Physics and Astronomy, Michigan State University

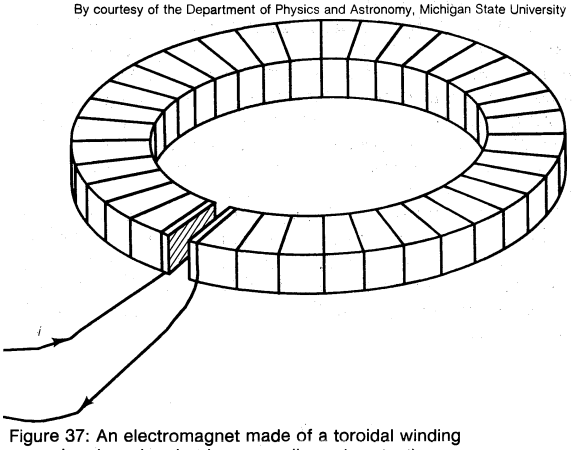


Figure 37: An electromagnet made of a toroidal winding around an iron ring that has a small gap (see text).

Because Figure 38 is a two-dimensional representation of a three-dimensional field, the spacing between the lines reflects the strength of the field only qualitatively. The high values of B between the two loops of the figure show that there is a large energy density in that region and separating the loops would reduce the energy. As discussed above, this is one more way of looking at the source of repulsion between these two loops. Figure 39 shows the B field for two loops with currents in the same direction. The force between the loops is attractive, and the distance separating them is equal to the loop radius. The result

By courtesy of the Department of Physics and Astronomy, Michigan State University

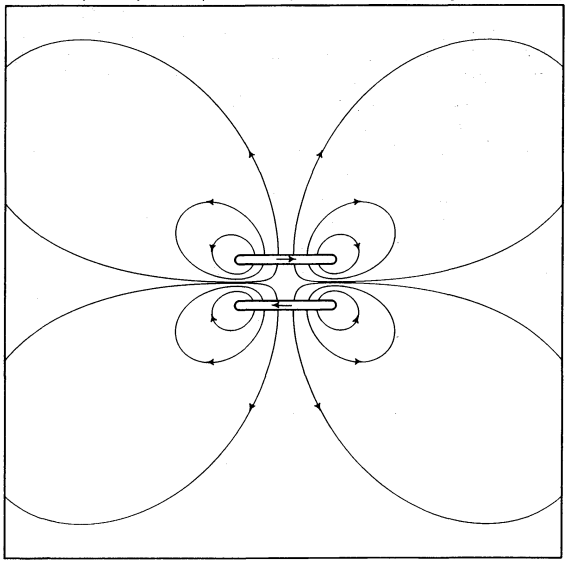


Figure 38: Magnetic field B of two current loops with currents in opposite directions (see text).

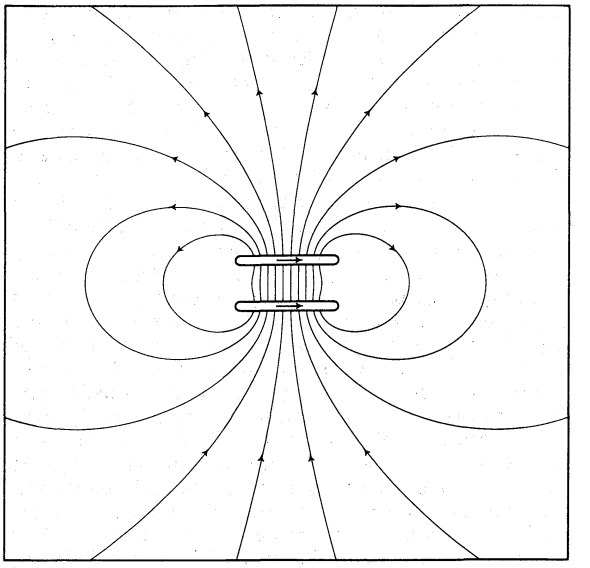


Figure 39: Magnetic field B of two current loops with currents in the same direction (see text).

By courtesy of the Department of Physics and Astronomy, Michigan State University

is that the B field in the central region between the two loops is homogeneous to a remarkably high degree. Such a configuration is called a Helmholtz coil. By carefully orienting and adjusting the current in a large Helmholtz coil, it is often possible to cancel an external magnetic field (such as the magnetic field of the Earth) in a region of space where experiments require the absence of all external magnetic fields.

Electromagnetism

The merger of electricity and magnetism from distinct phenomena into electromagnetism is tied to three closely related events. The first was Hans Christian Ørsted's accidental discovery of the influence of an electric current on a magnetic needle—namely, that magnetic fields are produced by electric currents. Ørsted's 1820 report of his observation spurred an intense effort by scientists to prove that magnetic fields can induce currents. The second event was Michael Faraday's experimental proof that a changing magnetic field can induce a current in a circuit. The third was James Clerk Maxwell's prediction that a changing electric field has an associated magnetic field. The technological revolution attributed to the development of electric power and radio communications can be traced to these three landmarks (see below).

EFFECTS OF VARYING MAGNETIC FIELDS

Faraday's law of induction. Faraday's discovery in 1831 of the phenomenon of magnetic induction is one of the great milestones in the quest toward understanding and exploiting nature. Stated simply, Faraday found that (1) a changing magnetic field in a circuit induces an electromotive force in the circuit; and (2) the magnitude of the electromotive force equals the rate at which the flux of the magnetic field through the circuit changes. The flux is a measure of how much field penetrates through the circuit. The electromotive force is measured in volts and is represented by the equation

Magnetic flux

$$\text{emf} = -\frac{d\Phi}{dt}. \qquad (43)$$

Here, Φ, the flux of the vector field B through the circuit, measures how much of the field passes through the circuit. To illustrate the meaning of flux, imagine how much water from a steady rain will pass through a circular ring of area A. When the ring is placed parallel to the path of the water drops, no water passes through the ring. The maximum rate at which drops of rain pass through the ring occurs when the surface is perpendicular to the motion of the drops. The rate of water drops crossing the

surface is the flux of the vector field $\rho \boldsymbol{v}$ through that surface, where ρ is the density of water drops and $\boldsymbol{v}$ represents the velocity of the water. Clearly, the angle between $\boldsymbol{v}$ and the surface is essential in determining the flux. To specify the orientation of the surface, a vector $\boldsymbol{A}$ is defined so that its magnitude is the surface area A in units of square metres and its direction is perpendicular to the surface. The rate at which raindrops pass through the surface is $\rho v \cos \theta A$, where θ is the angle between $\boldsymbol{v}$ and $\boldsymbol{A}$. Using vector notation, the flux is $\rho \boldsymbol{v} \cdot \boldsymbol{A}$. For the magnetic field, the amount of flux through a small area represented by the vector $d\boldsymbol{A}$ is given by $\boldsymbol{B} \cdot d\boldsymbol{A}$. For a circuit consisting of a single turn of wire, adding the contributions from the entire surface that is surrounded by the wire gives the magnetic flux Φ of equation (43). The rate of change of this flux is the induced electromotive force. The units of magnetic flux are webers, with one weber equaling one tesla per square metre. Finally, the minus sign in equation (43) indicates the direction of the induced electromotive force and hence of any induced current. The magnetic flux through the circuit generated by the induced current is in whatever direction will keep the total flux in the circuit from changing. The minus sign in equation (43) is an example of Lenz's law for magnetic systems. This law, deduced by the Russian-born physicist Heinrich Friedrich Emil Lenz, states that "what happens is that which opposes any change in the system."

Faraday's law is valid regardless of the process that causes the magnetic flux to change. It may be that a magnet is moved closer to a circuit or that a circuit is moved closer to a magnet. Figure 40 shows a magnet brought near a conducting ring and gives the direction of the induced current and field, thus illustrating both Faraday's and Lenz's laws. Another alternative is that the circuit may change in size in a fixed external magnetic field or, as in the case of alternating-current generation, that the circuit may be a coil of conducting wire rotating in a magnetic field so that the flux Φ varies sinusoidally in time.

By courtesy of the Department of Physics and Astronomy, Michigan State University

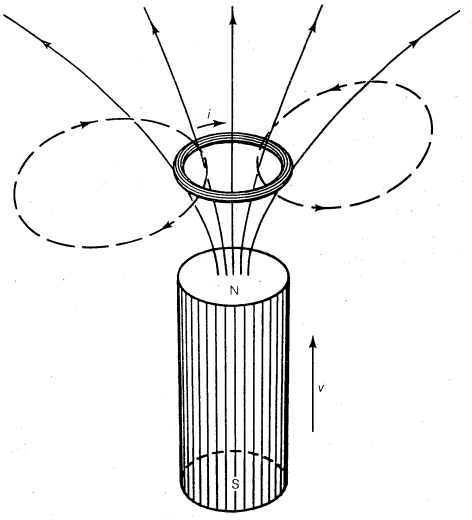


Figure 40: *Demonstration of Faraday's and Lenz's laws.* When a magnet is moved toward a conducting ring, an induced electromotive force causes a current i to flow in a direction such that the magnetic field inside the ring (represented by the two dashed field lines) opposes the increase of flux through the ring from the approaching magnet (see text).

The magnetic flux Φ through a circuit has to be considered carefully in the application of Faraday's law given in equation (43). For example, if a circuit consists of a coil with five closely spaced turns and if φ is the magnetic flux through a single turn, then the value of Φ for the five-turn circuit that must be used in Faraday's law is $\Phi = 5\varphi$. If the five turns are not the same size and closely spaced, the problem of determining Φ can be quite complex.

Self-inductance and mutual inductance. The self-inductance of a circuit is used to describe the reaction of the circuit to a changing current in the circuit, while the mutual inductance with respect to a second circuit describes the reaction to a changing current in the second circuit. When a current i_1 flows in circuit 1, i_1 produces a magnetic field $\boldsymbol{B}_1$; the magnetic flux through circuit 1 due to current i_1 is Φ_{11}. Since $\boldsymbol{B}_1$ is proportional to i_1, Φ_{11} is as well. The constant of proportionality is the self-inductance L_1 of the circuit. It is defined by the equation

$$\Phi_{11} = L_1 i_1. \tag{44}$$

As indicated earlier, the units of inductance are henrys. If a second circuit is present, some of the field $\boldsymbol{B}_1$ will pass through circuit 2 and there will be a magnetic flux Φ_{21} in circuit 2 due to the current i_1. The mutual inductance M_{21} is given by

$$\Phi_{21} = M_{21} i_1. \tag{45}$$

The magnetic flux in circuit 1 due to a current in circuit 2 is given by $\Phi_{12} = M_{12} i_2$. An important property of the mutual inductance is that $M_{21} = M_{12}$. It is therefore sufficient to use the label M without subscripts for the mutual inductance of two circuits.

The value of the mutual inductance of two circuits can range from $+\sqrt{L_1 L_2}$ to $-\sqrt{L_1 L_2}$, depending on the flux linkage between the circuits. If the two circuits are very far apart or if the field of one circuit provides no magnetic flux through the other circuit, the mutual inductance is zero. The maximum possible value of the mutual inductance of two circuits is approached as the two circuits produce $\boldsymbol{B}$ fields with increasingly similar spatial configurations.

If the rate of change with respect to time is taken for the terms on both sides of equation (44), the result is $d\Phi_{11}/dt = L_1 di_1/dt$. According to Faraday's law, $d\Phi_{11}/dt$ is the negative of the induced electromotive force. The result is the equation frequently used for a single inductor in an AC circuit—*i.e.*,

$$\text{emf} = -L\frac{di}{dt}. \tag{46}$$

The phenomenon of self-induction was first recognized by the American scientist Joseph Henry. He was able to generate large and spectacular electric arcs by interrupting the current in a large copper coil with many turns. While a steady current is flowing in a coil, the energy in the magnetic field is given by $\frac{1}{2}Li^2$. If both the inductance L and the current i are large, the amount of energy is also large. If the current is interrupted, as, for example, by opening a knife-blade switch, the current and therefore the magnetic flux through the coil drop quickly. Equation (46) describes the resulting electromotive force induced in the coil, and a large potential difference is developed between the two poles of the switch. The energy stored in the magnetic field of the coil is dissipated as heat and radiation in an electric arc across the space between the terminals of the switch. Due to advances in superconducting wires for electromagnets, it is possible to use large magnets with magnetic fields of several teslas for temporarily storing electric energy as energy in the magnetic field. This is done to accommodate short-term fluctuations in the consumption of electric power.

A transformer is an example of a device that uses circuits with maximum mutual induction. Figure 41 illustrates the configuration of a typical transformer. Here, coils of insulated conducting wire are wound around a ring of iron constructed of thin isolated laminations or sheets. The laminations minimize eddy currents in the iron. Eddy

By courtesy of the Department of Physics and Astronomy, Michigan State University

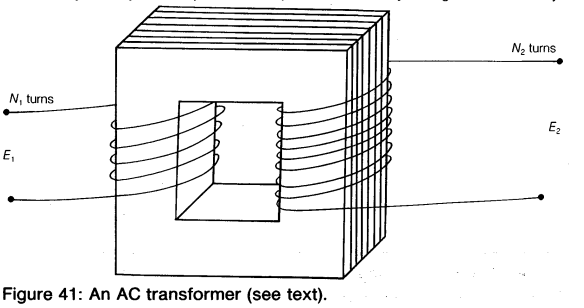


Figure 41: An AC transformer (see text).

Eddy currents

currents are circulatory currents induced in the metal by the changing magnetic field. These currents produce an undesirable by-product—heat in the iron. Energy loss in a transformer can be reduced by using thinner laminations, very "soft" (low-carbon) iron and wire with a larger cross section, or by winding the primary and secondary circuits with conductors that have very low resistance. Unfortunately, reducing the heat loss increases the cost of transformers. Transformers used to transmit and distribute power are commonly 98 to 99 percent efficient. While eddy currents are a problem in transformers, they are useful for heating objects in a vacuum. Eddy currents are induced in the object to be heated by surrounding a relatively nonconducting vacuum enclosure with a coil carrying a high-frequency alternating current.

In a transformer, the iron ensures that nearly all the lines of $\boldsymbol{B}$ passing through one circuit also pass through the second circuit and that, in fact, essentially all the magnetic flux is confined to the iron. Each turn of the conducting coils has the same magnetic flux; thus, the total flux for each coil is proportional to the number of turns in the coil. As a result, if a source of sinusoidally varying electromotive force is connected to one coil, the electromotive force in the second coil is given by

$$\text{emf}_2 = \text{emf}_1 \frac{N_2}{N_1}. \tag{47}$$

Thus, depending on the ratio of N_2 to N_1, the transformer can be either a step-up or a step-down device for alternating voltages. For many reasons, including safety, generation and consumption of electric power occur at relatively low voltages. Step-up transformers are used to obtain high voltages before electric power is transmitted, since for a given amount of power, the current in the transmission lines is much smaller. This minimizes energy lost by resistive heating of the conductors.

Faraday's law constitutes the basis for the power industry and for the transformation of mechanical energy into electric energy. In 1821, a decade before his discovery of magnetic induction, Faraday conducted experiments with electric wires rotating around compass needles. This earlier work, in which a wire carrying a current rotated around a magnetized needle and a magnetic needle was made to rotate around a wire carrying an electric current, provided the groundwork for the development of the electric motor.

EFFECTS OF VARYING ELECTRIC FIELDS

Maxwell's prediction that a changing electric field generates a magnetic field was a masterstroke of pure theory. The Maxwell equations for the electromagnetic field unified all that was hitherto known about electricity and magnetism and predicted the existence of an electromagnetic phenomenon that can travel as waves with the velocity of $1/\sqrt{\varepsilon_0\mu_0}$ in a vacuum. That velocity, which is based on constants obtained from purely electric measurements, corresponds to the speed of light. Consequently, Maxwell concluded that light itself was an electromagnetic phenomenon. Later, Einstein's special relativity theory postulated that the value of the speed of light is independent of the motion of the source of the light. Since then, the speed of light has been measured with increasing accuracy. In 1983 it was defined to be exactly 299,792,458 metres per second. Together with the cesium clock, which has been used to define the second, the speed of light serves as the new standard for length.

The circuit in Figure 42 is an example of a magnetic field generated by a changing electric field. A capacitor with parallel plates is charged at a constant rate by a steady current flowing through the long, straight leads in Figure 42A.

Ampère's law in circuital form

The objective is to apply Ampère's circuital law for magnetic fields to the path P, which goes around the wire in Figure 42A. This law (named in honour of the French physicist André-Marie Ampère) can be derived from the Biot and Savart equation for the magnetic field produced by a current (equation [34]). Using vector calculus notation, Ampère's law states that the integral $\oint \boldsymbol{B} \cdot d\boldsymbol{l}$ along a closed path surrounding the current i is equal to $\mu_0 i$. (An integral is essentially a sum, and, in this case, $\oint \boldsymbol{B} \cdot d\boldsymbol{l}$ is the sum of $B \cos \theta dl$ taken for a small length of the path until the complete loop is included. At each segment of the path dl, θ is the angle between the field $\boldsymbol{B}$ and $d\boldsymbol{l}$.) The current i in Ampère's law is the total flux of the current density $\boldsymbol{J}$ through any surface surrounded by the closed path. In Figure 42A, the closed path is labeled P, and a surface S_1 is surrounded by path P. All the current density through S_1 lies within the conducting wire. The total flux of the current density is the current i flowing through the wire. The result for surface S_1 reflects the value of the magnetic field around the wire in the region of the path P. In Figure 42B, path P is the same but the surface S_2 passes between the two plates of the capacitor. The value of the total flux of the current density through the surface should also be i. There is, however, clearly no motion of charge at all through the surface S_2. The dilemma is that the value of the integral $\oint \boldsymbol{B} \cdot d\boldsymbol{l}$ for the path P cannot be both $\mu_0 i$ and zero.

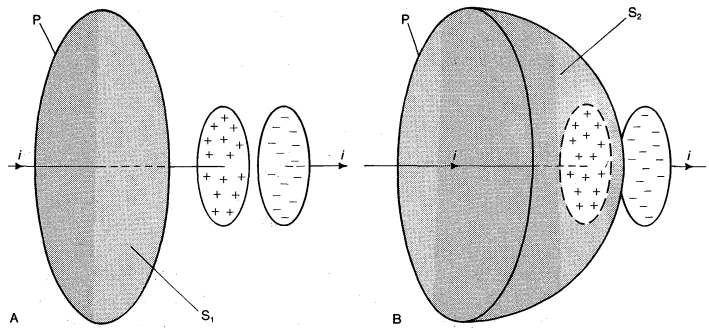


Figure 42: Current i charging a capacitor as an illustration of Maxwell's displacement current (see text).
By courtesy of the Department of Physics and Astronomy, Michigan State University

Maxwell's resolution of this dilemma was his conclusion that there must be some other kind of current density, called the displacement current $\boldsymbol{J}_d$, for which the total flux through the surface S_2 would be the same as the current i through the surface S_1. $\boldsymbol{J}_d$ would take, for the surface S_2, the place of the current density $\boldsymbol{J}$ associated with the movement of charge, since $\boldsymbol{J}$ is clearly zero due to the lack of charges between the plates of the capacitor. What happens between the plates while the current i is flowing? Because the amount of charge on the capacitor increases with time, the electric field between the plates increases with time too. If the current stops, there is an electric field between the plates as long as the plates are charged, but there is no magnetic field around the wire. Maxwell decided that the new type of current density was associated with the changing of the electric field. He found that

$$\boldsymbol{J}_d = \frac{d\boldsymbol{D}}{dt}, \tag{48}$$

where $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$ and $\boldsymbol{E}$ is the electric field between the plates. In situations where matter is present, the field $\boldsymbol{D}$ in equation (48) is modified to include polarization effects; the result is $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}$. The field $\boldsymbol{D}$ is measured in coulombs per square metre. Adding the displacement current to Ampère's law represented Maxwell's prediction that a changing electric field also could be a source of the magnetic field $\boldsymbol{B}$. Following Maxwell's predictions of electromagnetic waves, the German physicist Heinrich Hertz initiated the era of radio communications in 1887 by generating and detecting electromagnetic waves.

Maxwell's equations

Using vector calculus notation, the four equations of Maxwell's theory of electromagnetism are

$$\text{I. div } \boldsymbol{D} = \rho, \tag{49}$$

$$\text{II. div } \boldsymbol{B} = 0, \tag{50}$$

$$\text{III. } \textit{curl } \boldsymbol{E} = -\frac{d\boldsymbol{B}}{dt}, \tag{51}$$

$$\text{IV. } \textit{curl } \boldsymbol{H} = \boldsymbol{J} + \frac{d\boldsymbol{D}}{dt}, \tag{52}$$

where $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}$, and $\boldsymbol{H} = \boldsymbol{B}/\mu_0 - \boldsymbol{M}$. The first equation is based on Coulomb's inverse square law for the force between two charges; it is a form of Gauss's law, which

relates the flux of the electric field through a closed surface to the total charge enclosed by the surface. The second equation is based on the fact that apparently no magnetic monopoles exist in nature; if they did, they would be point sources of magnetic field. The third is a statement of Faraday's law of magnetic induction, which reveals that a changing magnetic field generates an electric field. The fourth is Ampère's law as extended by Maxwell to include the displacement current discussed above; it associates a magnetic field to a changing electric field as well as to an electric current.

Maxwell's four equations represent a complete description of the classical theory of electromagnetism. His discovery that light is an electromagnetic wave meant that optics could be understood as part of electromagnetism. Only in microscopic situations is it necessary to modify Maxwell's equations to include quantum effects. That modification, known as quantum electrodynamics (QED), accounts for certain atomic properties to a degree of precision exceeding one part in 100 million.

Sometimes it is necessary to shield apparatus from external electromagnetic fields. For a static electric field, this is a simple matter; the apparatus is surrounded by a shield made of a good conductor (*e.g.*, copper). Shielding apparatus from a steady magnetic field is more difficult because materials with infinite magnetic permeability μ do not exist; for example, a hollow shield made of soft iron will reduce the magnetic field inside to a considerable extent but not completely. As discussed earlier, it is sometimes possible to superpose a field in the opposite direction to produce a very low field region and then to use additional material with a high μ for shielding. In the case of electromagnetic waves, the penetration of the waves in matter varies, depending on the frequency of the radiation and the electric conductivity of the medium. The skin depth δ (which is the distance in the conducting medium traversed for an amplitude decrease of $1/e$, about 1/3) is given by

$$\delta = \sqrt{\frac{2}{\omega \mu_0 \sigma_J}}.$$

At high frequency, the skin depth is small. Therefore, to transmit electronic messages through seawater, for example, a very low frequency must be used to get a reasonable fraction of the signal far below the surface.

A metal shield can have some holes in it and still be effective. For instance, a typical microwave oven has a frequency of 2.5 gigahertz, which corresponds to a wavelength of about 12 centimetres for the electromagnetic wave inside the oven. The metal shield on the door has small holes about two millimetres in diameter; the shield works because the wavelength of the microwave radiation is much greater than the size of the holes. On the other hand, the same shield is not effective with radiation of a much shorter wavelength. Visible light passes through the holes in the shield, as evidenced by the fact that it is possible to see inside a microwave oven when the door is closed. (E.Ka./S.McG.)

Electric properties of matter

PIEZOELECTRICITY

Some solids, notably certain crystals, have permanent electric polarization. Other crystals become electrically polarized when subjected to stress. In electric polarization, the centre of positive charge within an atom, molecule, or crystal lattice element is separated slightly from the centre of negative charge. Piezoelectricity (literally "pressure electricity") is observed if a stress is applied to a solid, for example, by bending, twisting, or squeezing it. If a thin slice of quartz is compressed between two electrodes, a potential difference occurs; conversely, if the quartz crystal is inserted into an electric field, the resulting stress changes its dimensions. Piezoelectricity is responsible for the great precision of clocks and watches equipped with quartz oscillators. It also is used in electric guitars and various other musical instruments to transform mechanical vibrations into corresponding electric signals, which are then amplified and converted to sound by acoustical speakers.

Electric polarization due to mechanical stress

A crystal under stress exhibits the direct piezoelectric effect; a polarization P, proportional to the stress, is produced. In the converse effect, an applied electric field produces a distortion of the crystal, represented by a strain proportional to the applied field. The basic equations of piezoelectricity are $P = d \times stress$ and $E = strain/d$. The piezoelectric coefficient d (in metres per volt) is approximately 3×10^{-12} for quartz, 5×-10^{-11} for ammonium dihydrogen phosphate, and 3×10^{-10} for lead zirconate titanate.

For an elastic body, the stress is proportional to the strain—*i.e.*, $stress = Y_e \times strain$. The proportionality constant is the coefficient of elasticity Y_e, also called Young's modulus for the English physicist Thomas Young. Using that relation, the induced polarization can be written as $P = dY_e \times strain$, while the stress required to keep the strain constant when the crystal is in an electric field is $stress = -dY_eE$. The strain in a deformed elastic body is the fractional change in the dimensions of the body in various directions; the stress is the internal pressure along the various directions. Both are second-rank tensors, and, since electric field and polarization are vectors, the detailed treatment of piezoelectricity is complex. The equations above are oversimplified but can be used for crystals in certain orientations.

The polarization effects responsible for piezoelectricity arise from small displacements of ions in the crystal lattice. Such an effect is not found in crystals with a centre of symmetry. The direct effect can be quite strong; a potential $V = Y_e d\delta/\varepsilon_0 K$ is generated in a crystal compressed by an amount δ, where K is the dielectric constant. If lead zirconate titanate is placed between two electrodes and a pressure causing a reduction of only 1/20th of one millimetre is applied, a 100,000-volt potential is produced. The direct effect is used, for example, to generate an electric spark with which to ignite natural gas in a heating unit or an outdoor cooking grill.

In practice, the converse piezoelectric effect, which occurs when an external electric field changes the dimensions of a crystal, is small because the electric fields that can be generated in a laboratory are minuscule compared to those existing naturally in matter. A static electric field of 10^6 volts per metre produces a change of only about 0.001 millimetre in the length of a one-centimetre quartz crystal. The effect can be enhanced by the application of an alternating electric field of the same frequency as the natural mechanical vibration frequency of the crystal. Many of the crystals have a quality factor Q of several hundred, and, in the case of quartz, the value can be 10^6. The result is a piezoelectric coefficient a factor Q higher than for a static electric field. The very large Q of quartz is exploited in electronic oscillator circuits to make remarkably accurate timepieces. The mechanical vibrations that can be induced in a crystal by the converse piezoelectric effect are also used to generate ultrasound, which is sound with a frequency far higher than frequencies audible to the human ear—above 20 kilohertz. The reflected sound is detectable by the direct effect. Such effects form the basis of ultrasound systems used to fathom the depths of lakes and waterways and to locate fish. Ultrasound has found application in medical imaging (*e.g.*, fetal monitoring and the detection of abnormalities such as prostate tumours). The use of ultrasound makes it possible to produce detailed pictures of organs and other internal structures because of the variation in the reflection of sound from various body tissues. Thin films of polymeric plastic with a piezoelectric coefficient of about 10^{-11} metres per volt are being developed and have numerous potential applications as pressure transducers.

Generation of ultrasound

ELECTRO-OPTIC PHENOMENA

The index of refraction n of a transparent substance is related to its electric polarizability and is given by $n^2 = 1 + \chi_e/\varepsilon_0$. As discussed earlier, χ_e is the electric susceptibility of a medium, and the equation $\boldsymbol{P} = \chi_e \boldsymbol{E}$ relates the polarization of the medium to the applied electric field. For most matter, χ_e is not a constant independent of the value of the electric field, but rather depends to a small degree on the value of the field. Thus, the index of

refraction can be changed by applying an external electric field to a medium. In liquids, glasses, and crystals that have a centre of symmetry, the change is usually very small. Called the Kerr effect (for its discoverer, the Scottish physicist John Kerr), it is proportional to the square of the applied electric field. In noncentrosymmetric crystals, the change in the index of refraction n is generally much greater; it depends linearly on the applied electric field and is known as the Pockels effect (after the German physicist F. R. Pockels).

Modulation of the index of refraction

A varying electric field applied to a medium will modulate its index of refraction. This change in the index of refraction can be used to modulate light and make it carry information. A crystal widely used for its Pockels effect is potassium dihydrogen phosphate, which has good optical properties and low dielectric losses even at microwave frequencies.

An unusually large Kerr effect is found in nitrobenzene, a liquid with highly "acentric" molecules that have large electric dipole moments. Applying an external electric field partially aligns the otherwise randomly oriented dipole moments and greatly enhances the influence of the field on the index of refraction. The length of the path of light through nitrobenzene can be adjusted easily because it is a liquid.

THERMOELECTRICITY

When two metals are placed in electric contact, electrons flow out of the one in which the electrons are less bound and into the other. The binding is measured by the location of the so-called Fermi level of electrons in the metal; the higher the level, the lower is the binding. The Fermi level represents the demarcation in energy within the conduction band of a metal between the energy levels occupied by electrons and those that are unoccupied. The energy of an electron at the Fermi level is $-W$ relative to a free electron outside the metal. The flow of electrons between the two conductors in contact continues until the change in electrostatic potential brings the Fermi levels of the two metals (W_1 and W_2) to the same value. This electrostatic potential is called the contact potential φ_{12} and is given by $e\varphi_{12} = W_1 - W_2$, where e is 1.6×10^{-19} coulomb.

If a closed circuit is made of two different metals, there will be no net electromotive force in the circuit because the two contact potentials oppose each other and no current will flow. There will be a current if the temperature of one of the junctions is raised with respect to that of the second.

Generation of a thermal electromotive force

There is a net electromotive force generated in the circuit, as it is unlikely that the two metals will have Fermi levels with identical temperature dependence. To maintain the temperature difference, heat must enter the hot junction and leave the cold junction; this is consistent with the fact that the current can be used to do mechanical work. The generation of a thermal electromotive force at a junction is called the Seebeck effect (after the Estonian-born German physicist Thomas Johann Seebeck). The electromotive force is approximately linear with the temperature difference between two junctions of dissimilar metals, which are called a thermocouple. For a thermocouple made of iron and constantan (an alloy of 60 percent copper and 40 percent nickel), the electromotive force is about five millivolts when the cold junction is at 0° C and the hot junction at 100° C. One of the principal applications of the Seebeck effect is the measurement of temperature. The chemical properties of the medium, the temperature of which is measured, and the sensitivity required dictate the choice of components of a thermocouple.

The absorption or release of heat at a junction in which there is an electric current is called the Peltier effect (after the French physicist Jean-Charles Peltier). Both the Seebeck and Peltier effects also occur at the junction between a metal and a semiconductor and at the junction between two semiconductors. The development of semiconductor thermocouples (*e.g.*, those consisting of *n*-type and *p*-type bismuth telluride) has made the use of the Peltier effect practical for refrigeration. Sets of such thermocouples are connected electrically in series and thermally in parallel. When an electric current is made to flow, a temperature difference, which depends on the current, develops between the two junctions. If the temperature of the hotter junction is kept low by removing heat, the second junction can be tens of degrees colder and act as a refrigerator. Peltier refrigerators are used to cool small bodies; they are compact, have no moving mechanical parts, and can be regulated to maintain precise and stable temperatures. They are employed in numerous applications, as, for example, to keep the temperature of a sample constant while it is on a microscope stage.

THERMIONIC EMISSION

A metal contains mobile electrons in a partially filled band of energy levels—*i.e.*, the conduction band. These electrons, though mobile within the metal, are rather tightly bound to it. The energy that is required to release a mobile electron from the metal varies from about 1.5 to approximately six electron volts, depending on the metal. In thermionic emission, some of the electrons acquire enough energy from thermal collisions to escape from the metal. The number of electrons emitted and therefore the thermionic emission current depend critically on temperature.

In a metal the conduction-band levels are filled up to the Fermi level, which lies at an energy $-W$ relative to a free electron outside the metal. The work function of the metal, which is the energy required to remove an electron from the metal, is therefore equal to W. At a temperature of 1,000 K only a small fraction of the mobile electrons have sufficient energy to escape. The electrons that can escape are moving so fast in the metal and have such high kinetic energies that they are unaffected by the periodic potential caused by atoms of the metallic lattice. They behave like electrons trapped in a region of constant potential. Because of this, when the rate at which electrons escape from the metal is calculated, the detailed structure of the metal has little influence on the final result. A formula known as Richardson's law (first proposed by the English physicist Owen W. Richardson) is roughly valid for all metals. It is usually expressed in terms of the emission current density (J) as

$$J = AT^2e^{-W/kT}$$

in amperes per square metre. The Boltzmann constant k has the value 8.62×10^{-5} electron volts per kelvin, and temperature T is in kelvins. The constant A is 1.2×10^6 ampere degree squared per square metre, and varies slightly for different metals. For tungsten, which has a work function W of 4.5 electron volts, the value of A is 7×10^5 amperes per square metre kelvin squared and the current density at T equaling 2,400 K is 0.14 ampere per square centimetre. J rises rapidly with temperature. If T is increased to 2,600 K, J rises to 0.9 ampere per square centimetre. Tungsten does not emit appreciably at 2,000 K or below (less than 0.05 milliampere per square centimetre) because its work function of 4.5 electron volts is large compared to the thermal energy kT, which is only 0.16 electron volt. In vacuum tubes, the cathode usually is coated with a mixture of barium and strontium oxides. At 1,000 K the oxide has a work function of approximately 1.3 electron volts and is a reasonably good conductor. Currents of several amperes per square centimetre can be drawn from oxide cathodes, but in practice the current density is generally less than 0.2 ampere per square centimetre. The oxide layer deteriorates rapidly when higher current densities are drawn.

SECONDARY ELECTRON EMISSION

If electrons with energies of 10 to 1,000 electron volts strike a metal surface in a vacuum, their energy is lost in collisions in a region near the surface, and most of it is transferred to other electrons in the metal. Because this occurs near the surface, some of these electrons may be ejected from the metal and form a secondary emission current. The ratio of secondary electrons to incident electrons is known as the secondary emission coefficient. For low-incident energies (below about one electron volt), the primary electrons tend to be reflected and the secondary emission coefficient is near unity. With increasing energy, the coefficient at first falls and then at about 10 elec-

tron volts begins to rise again, usually reaching a peak of value between 2 and 4 at energies of a few hundred electron volts. At higher energies, the primary electrons penetrate so far below the surface before losing energy that the excited electrons have little chance of reaching the surface and escaping. The secondary emission coefficients fall and, when the electrons have energies exceeding 20 kiloelectron volts, are usually well below unity. Secondary emission also can occur in insulators. Because many insulators have rather high secondary emission coefficients, it is often useful when high secondary emission yields are required to coat a metal electrode with a thin insulator layer a few atoms thick.

PHOTOELECTRIC CONDUCTIVITY

If light with a photon energy hv that exceeds the work function W falls on a metal surface, some of the incident photons will transfer their energy to electrons, which then will be ejected from the metal. Since hv is greater than W, the excess energy $hv - W$ transferred to the electrons will be observed as their kinetic energy outside the metal. The relation between electron kinetic energy E and the frequency v (that is, $E = hv - W$) is known as the Einstein relation, and its experimental verification helped to establish the validity of quantum theory. The energy of the electrons depends on the frequency of the light, while the intensity of the light determines the rate of photoelectric emission.

In a semiconductor the valence band of energy levels is almost completely full while the conduction band is almost empty. The conductivity of the material derives from the few holes present in the valence band and the few electrons in the conduction band. Electrons can be excited from the valence to the conduction band by light photons having an energy hv that is larger than energy gap E_g between the bands. The process is an internal photoelectric effect. The value of E_g varies from semiconductor to semiconductor. For lead sulfide, the threshold frequency occurs in the infrared, whereas for zinc oxide it is in the ultraviolet. For silicon, E_g equals 1.1 electron volts, and the threshold wavelength is in the infrared, about 1,100 nanometres. Visible radiation produces electron transitions with almost unity quantum efficiency in silicon. Each transition yields a hole–electron pair (*i.e.,* two carriers) that contributes to electric conductivity. For example, if one milliwatt of light strikes a sample of pure silicon in the form of a thin plate one square centimetre in area and 0.03 centimetre thick (which is thick enough to absorb all incident light), the resistance of the plate will be decreased by a factor of about 1,000. In practice, photoconductive effects are not usually as large as this, but this example indicates that appreciable changes in conductivity can occur even with low illumination. Photoconductive devices are simple to construct and are used to detect visible, infrared, and ultraviolet radiation.

Internal photoelectric effect

ELECTROLUMINESCENCE

Conduction electrons moving in a solid under the influence of an electric field usually lose kinetic energy in low-energy collisions as fast as they acquire it from the field. Under certain circumstances in semiconductors, however, they can acquire enough energy between collisions to excite atoms in the next collision and produce radiation as the atoms de-excite. A voltage applied across a thin layer of zinc sulfide powder causes just such an electroluminescent effect. Electroluminescent panels are of more interest as signal indicators and display devices than as a source of general illumination.

A somewhat similar effect occurs at the junction in a reverse-biased semiconductor p–n junction diode—*i.e.,* a p–n junction diode in which the applied potential is in the direction of small current flow. Electrons in the intense field at the depleted junction easily acquire enough energy to excite atoms. Little of this energy finally emerges as light, though the effect is readily visible under a microscope.

When a junction between a heavily doped n-type material and a less doped p-type material is forward-biased so that a current will flow easily, the current consists mainly of electrons injected from the n-type material into the conduction band of the p-type material. These electrons ultimately drop into holes in the valence band and release energy equal to the energy gap of the material. In most cases, this energy E_g is dissipated as heat, but in gallium phosphide and especially in gallium arsenide, an appreciable fraction appears as radiation, the frequency v of which satisfies the relation $hv = E_g$. In gallium arsenide, though up to 30 percent of the input electric energy is available as radiation, the characteristic wavelength of 900 nanometres is in the infrared. Gallium phosphide gives off visible green light but is inefficient; other related III-V compound semiconductors emit light of different colours. Electroluminescent injection diodes of such materials, commonly known as light-emitting diodes (LEDs), are employed mainly as indicator lamps and numeric displays. Semiconductor lasers built with layers of indium phosphide and of gallium indium arsenide phosphide have proved more useful. Unlike gas or optically pumped lasers, these semiconductor lasers can be modulated directly at high frequencies. Not only are they used in devices such as compact digital disc players but also as light sources for long-distance optical fibre communications systems (see ELECTRONICS: *Light-emitting diodes and semiconductor lasers*).

BIOELECTRIC EFFECTS

Bioelectricity refers to the generation or action of electric currents or voltages in biological processes. Bioelectric phenomena include fast signaling in nerves and the triggering of physical processes in muscles or glands. There is some similarity among the nerves, muscles, and glands of all organisms, possibly because fairly efficient electrochemical systems evolved early. Scientific studies tend to focus on the following: nerve or muscle tissue; such organs as the heart, brain, eye, ear, stomach, and certain glands; electric organs in some fish; and potentials associated with damaged tissue.

Electric activity in living tissue is a cellular phenomenon, dependent on the cell membrane. The membrane acts like a capacitor, storing energy as electrically charged ions on opposite sides of the membrane. The stored energy is available for rapid utilization and stabilizes the membrane system so that it is not activated by small disturbances.

Cells capable of electric activity show a resting potential in which their interiors are negative by about 0.1 volt or less compared with the outside of the cell. When the cell is activated, the resting potential may reverse suddenly in sign; as a result, the outside of the cell becomes negative and the inside positive. This condition lasts for a short time, after which the cell returns to its original resting state. This sequence, called depolarization and repolarization, is accompanied by a flow of substantial current through the active cell membrane, so that a "dipole-current source" exists for a short period. Small currents flow from this source through the aqueous medium containing the cell and are detectable at considerable distances from it. These currents, originating in active membrane, are functionally significant very close to their site of origin but must be considered incidental at any distance from it. In electric fish, however, adaptations have occurred, and this otherwise incidental electric current is actually utilized. In some species the external current is apparently used for sensing purposes, while in others it is used to stun or kill prey. In both cases, voltages from many cells add up in series, thus assuring that the specialized functions can be performed. Bioelectric potentials detected at some distance from the cells generating them may be as small as the 20 or 30 microvolts associated with certain components of the human electroencephalogram or the millivolt of the human electrocardiogram. On the other hand, electric eels can deliver electric shocks with voltages as large as 1,000 volts.

Depolarization and repolarization in cells

In addition to the potentials originating in nerve or muscle cells, relatively steady or slowly varying potentials (often designated dc) are known. These dc potentials occur in the following cases: in areas where cells have been damaged and where ionized potassium is leaking (as much as 50 millivolts); when one part of the brain is compared with another part (up to one millivolt); when different areas of the skin are compared (up to 10 millivolts); within pockets in active glands, *e.g.,* follicles in the thyroid (as

high as 60 millivolts); and in special structures in the inner ear (about 80 millivolts).

A small electric shock caused by static electricity during cold, dry weather is a familiar experience. While the sudden muscular reaction it engenders is sometimes unpleasant, it is usually harmless. Even though static potentials of several thousand volts are involved, a current exists for only a brief time and the total charge is very small. A steady current of two milliamperes through the body is barely noticeable. Severe electrical shock can occur above 10 milliamperes, however. Lethal current levels range from 100 to 200 milliamperes. Larger currents, which produce burns and unconsciousness, are not fatal if the victim is given prompt medical care. (Above 200 milliamperes, the heart is clamped during the shock and does not undergo ventricular fibrillation.) Prevention clearly includes avoiding contact with live electric wiring; risk of injury increases considerably if the skin is wet, as the electric resistance of wet skin may be hundreds of times smaller than that of dry skin. (F.N.H.R./E.E.S./E.Ka.)

Magnetic properties of matter

All matter exhibits magnetic properties when placed in an external magnetic field. Even substances like copper and aluminum that are not normally thought of as having magnetic properties are affected by the presence of a magnetic field such as that produced by either pole of a bar magnet. Depending on whether there is an attraction or repulsion by the pole of a magnet, matter is classified as being either paramagnetic or diamagnetic, respectively. A few materials, notably iron, show a very large attraction toward the pole of a permanent bar magnet; materials of this kind are called ferromagnetic.

Classification of matter as paramagnetic, diamagnetic, or ferromagnetic

In 1845 Faraday became the first to classify substances as either diamagnetic or paramagnetic. He based this classification on his observation of the force exerted on substances in an inhomogeneous magnetic field. At moderate field strengths, the magnetization $\boldsymbol{M}$ of a substance is linearly proportional to the strength of the applied field $\boldsymbol{H}$. The magnetization is specified by the magnetic susceptibility χ (previously labeled χ_m), defined by the relation $\boldsymbol{M} = \chi \boldsymbol{H}$. A sample of volume V placed in a field $\boldsymbol{H}$ directed in the x-direction and increasing in that direction at a rate $d\boldsymbol{H}/dx$ will experience a force in the x-direction of $\boldsymbol{F} = \chi \mu_0 V H\ (d\boldsymbol{H}/dx)$. If the magnetic susceptibility χ is positive, the force is in the direction of increasing field strength, whereas if χ is negative, it is in the direction of decreasing field strength. Measurement of the force $\boldsymbol{F}$ in a known field $\boldsymbol{H}$ with a known gradient $d\boldsymbol{H}/dx$ is the basis of a number of accurate methods of determining χ.

Substances for which the magnetic susceptibility is negative (*e.g.*, copper and silver) are classified as diamagnetic. The susceptibility is small, on the order of -10^{-5} for solids and liquids and -10^{-8} for gases. A characteristic feature of diamagnetism is that the magnetic moment per unit mass in a given field is virtually constant for a given substance over a very wide range of temperatures. It changes little between solid, liquid, and gas; the variation in the susceptibility between solid or liquid and gas is almost entirely due to the change in the number of molecules per unit volume. This indicates that the magnetic moment induced in each molecule by a given field is primarily a property characteristic of the molecule.

Substances for which the magnetic susceptibility is positive are classed as paramagnetic. In a few cases (including most metals), the susceptibility is independent of temperature, but in most compounds it is strongly temperature dependent, increasing as the temperature is lowered. Measurements by the French physicist Pierre Curie in 1895 showed that for many substances the susceptibility is inversely proportional to the absolute temperature T; that is, $\chi = C/T$. This approximate relationship is known as Curie's law and the constant C as the Curie constant. A more accurate equation is obtained in many cases by modifying the above equation to $\chi = C/(T - \theta)$, where θ is a constant. This equation is called the Curie–Weiss law (after Curie and Pierre-Ernest Weiss, another French physicist). From the form of this last equation, it is clear that at the temperature $T = \theta$, the value of the susceptibility becomes infinite. Below this temperature, the material exhibits spontaneous magnetization—*i.e.*, it becomes ferromagnetic. Its magnetic properties are then very different from those in the paramagnetic or high-temperature phase. In particular, although its magnetic moment can be changed by the application of a magnetic field, the value of the moment attained in a given field is not always the same; it depends on the previous magnetic, thermal, and mechanical treatment of the sample.

Curie–Weiss law

INDUCED AND PERMANENT ATOMIC MAGNETIC DIPOLES

Whether a substance is paramagnetic or diamagnetic is determined primarily by the presence or absence of free magnetic dipole moments (*i.e.*, those free to rotate) in its constituent atoms. When there are no free moments, the magnetization is produced by currents of the electrons in their atomic orbits. The substance is then diamagnetic, with a negative susceptibility independent of both field strength and temperature.

In matter with free magnetic dipole moments, the orientation of the moments is normally random and, as a result, the substance has no net magnetization. When a magnetic field is applied, the dipoles are no longer completely randomly oriented; more dipoles point with the field than against the field. When this results in a net positive magnetization in the direction of the field, the substance has a positive susceptibility and is classified as paramagnetic.

The forces opposing alignment of the dipoles with the external magnetic field are thermal in origin and thus weaker at low temperatures. The excess number of dipoles pointing with the field is determined by (mB/kT), where mB represents the magnetic energy and kT the thermal energy. When the magnetic energy is small compared to the thermal energy, the excess number of dipoles pointing with the field is proportional to the field and inversely proportional to the absolute temperature, corresponding to Curie's law. When the value of (mB/kT) is large enough to align nearly all the dipoles with the field, the magnetization approaches a saturation value.

There is a third category of matter in which intrinsic moments are not normally present but appear under the influence of an external magnetic field. The intrinsic moments of conduction electrons in metals behave this way. One finds a small positive susceptibility independent of temperature comparable with the diamagnetic contribution, so that the overall susceptibility of a metal may be positive or negative. The molar susceptibility of elements is shown in Figure 43.

In addition to the forces exerted on atomic dipoles by an external magnetic field, mutual forces exist between the dipoles. Such forces vary widely for different substances. Below a certain transition temperature depending on the substance, they produce an ordered arrangement of the orientations of the atomic dipoles even in the absence of an external field. The mutual forces tend to align neighbouring dipoles either parallel or antiparallel to one another. Parallel alignment of atomic dipoles throughout large volumes of the substance results in ferromagnetism, with a permanent magnetization on a macroscopic scale. On the other hand, if equal numbers of atomic dipoles are aligned in opposite directions and the dipoles are of the same size, there is no permanent macroscopic magnetization, and this is known as antiferromagnetism. If the atomic dipoles are of different magnitudes and those pointing in one direction are all different in size from those pointing in the opposite direction, there exists permanent magnetization on a macroscopic scale in an effect known as ferrimagnetism. A simple schematic representation of these different possibilities is shown in Figure 44.

Mutual forces between dipoles

In all cases, the material behaves as a paramagnet above the characteristic transition temperature; it acquires a macroscopic magnetic moment only when an external field is applied.

DIAMAGNETISM

When an electron moving in an atomic orbit is in a magnetic field $\boldsymbol{B}$, the force exerted on the electron pro-

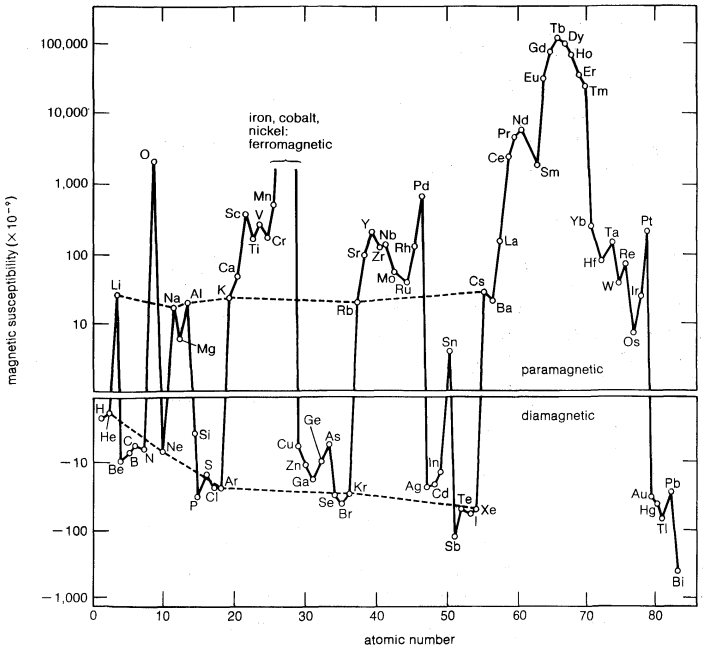


Figure 43: The susceptibility of a kilogram mole of the elements. Broken lines connect the alkali metals (paramagnetic) and the rare gases of the atmosphere (diamagnetic).

duces a small change in the orbital motion; the electron orbit precesses about the direction of $\boldsymbol{B}$. As a result, each electron acquires an additional angular momentum that contributes to the magnetization of the sample. The susceptibility χ is given by

$$\chi = -\mu_0 N \left(\frac{e^2}{6m}\right) \Sigma < r^2 >,$$

where $\Sigma < r^2 >$ is the sum of the mean square radii of all electron orbits in each atom, e and m are the charge and mass of the electron, and N is the number of atoms per unit volume. The negative sign of this susceptibility is a direct consequence of Lenz's law (see above). When $\boldsymbol{B}$ is switched on, the change in motion of each orbit is equivalent to an induced circulating electric current in such a direction that its own magnetic flux opposes the change in magnetic flux through the orbit; *i.e.*, the induced magnetic moment is directed opposite to $\boldsymbol{B}$.

Since the magnetization $\boldsymbol{M}$ is proportional to the number N of atoms per unit volume, it is sometimes useful to give the susceptibility per mole, χ_{mole}. For a kilogram mole (the molecular weight in kilograms), the numerical value of the molar susceptibility is

By courtesy of *Electrical Engineering*

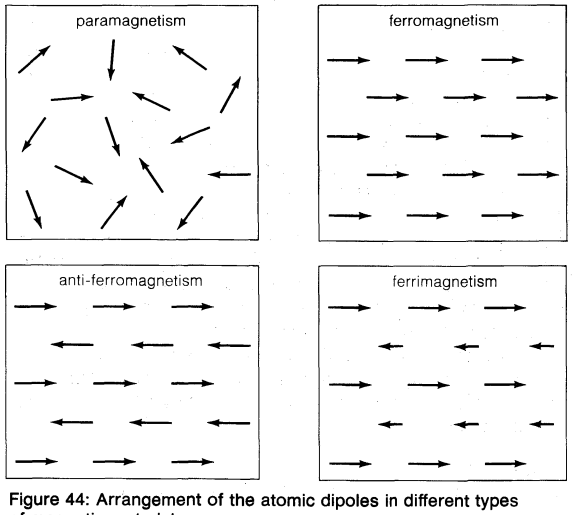


Figure 44: Arrangement of the atomic dipoles in different types of magnetic materials.

$$\chi_{mole} = -3.55 \times 10^{12} \Sigma < r^2 >.$$

For an atom, the mean value of $\Sigma < r^2 >$ is about 10^{-21} square metre and χ_{mole} has values of 10^{-9} to 10^{-10}; the atomic number Z equals the number of electrons in each atom. The quantity $\Sigma < r^2 >$ for each atom, and therefore the diamagnetic susceptibility, is essentially independent of temperature. It is also not affected by the surroundings of the atom.

Diamagnetism in superconductors

A different kind of diamagnetism occurs in superconductors. The conduction electrons are spread out over the entire metal, and so the induced magnetic moment is governed by the size of the superconducting sample rather than by the size of the individual constituent atoms (a very large effective $< r^2 >$). The diamagnetism is so strong that the magnetic field is kept out of the superconductor.

PARAMAGNETISM

Paramagnetism occurs primarily in substances in which some or all of the individual atoms, ions, or molecules possess a permanent magnetic dipole moment. The magnetization of such matter depends on the ratio of the magnetic energy of the individual dipoles to the thermal energy. This dependence can be calculated in quantum theory and is given by the Brillouin function, which depends only on the ratio (B/T). At low magnetic fields, the magnetization is linearly proportional to the field and reaches its maximum saturation value when the magnetic energy is much greater than the thermal energy. Figure 45 shows the dependence of the magnetic moment per ion in units of Bohr magnetons as a function of B/T. (One Bohr magneton equals 9.274×10^{-24} ampere times square metre.)

In substances that have a nuclear magnetic dipole moment, there is a further contribution to susceptibility. The size of the nuclear magnetic moment is only about one-thousandth that of an atom. Per kilogram mole, χ_n is on the order of $10^{-8}/T$; in solid hydrogen this just exceeds the electronic diamagnetism of 1 K.

Curie's law should hold when mB is much smaller than kT, provided that no other forces act on the atomic dipoles. In many solids, the presence of internal forces may cause the susceptibility to vary in a complicated way. If the forces orient the dipoles parallel to each other, the behaviour is ferromagnetic (see below). The forces may orient the dipoles so that the normal state has no free moment. If the force is sufficiently weak, a small magnetic field can reorient the dipoles, resulting in a net magnetization. This type of paramagnetism occurs for conduction electrons in a metal. In normal metals, each occupied electron state has two electrons with opposite spin orientation. This is a consequence of the Pauli principle of quantum mechanics, which permits no greater occupancy of the energetically favoured states. In the presence of a magnetic field, however, it is energetically more favourable for some of the electrons to move to higher states. With only single electrons in these states, the electron moments can be oriented along the field. The resulting paramagnetic susceptibility is independent of temperature. The net susceptibility is independent of temperature. The net susceptibility of a metal can be of either sign, since the diamagnetic and paramagnetic contributions are of comparable magnitudes.

FERROMAGNETISM

A ferromagnetic substance contains permanent atomic magnetic dipoles that are spontaneously oriented parallel to one another even in the absence of an external field. The magnetic repulsion between two dipoles aligned side by side with their moments in the same direction makes it difficult to understand the phenomenon of ferromagnetism. It is known that within a ferromagnetic material, there is a spontaneous alignment of atoms in large clusters. A new type of interaction, a quantum mechanical effect known as the exchange interaction, is involved. A highly simplified description of how the exchange interaction aligns electrons in ferromagnetic materials is given here.

Role of the exchange interaction

The magnetic properties of iron are thought to be the result of the magnetic moment associated with the spin

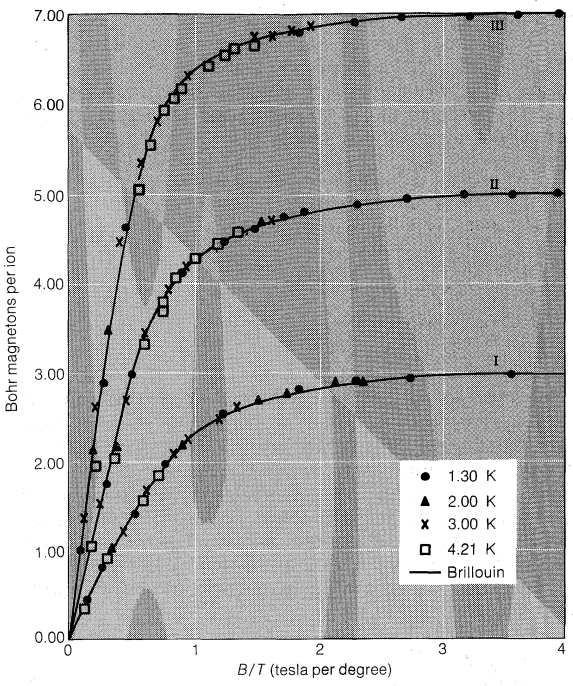


Figure 45: *The approach to saturation in the magnetization of a paramagnetic substance following a Brillouin curve.* The curves I, II, and III refer to ions of chromium, potassium alum, iron ammonium alum, and gadolinium sulfate octahydrate for which $g = 2$ and $j = 3/2$, 5/2, and 7/2, respectively.

of an electron in an outer atomic shell—specifically, the third *d* shell. Such electrons are referred to as magnetization electrons. The Pauli exclusion principle prohibits two electrons from having identical properties; for example, no two electrons can be in the same location and have spins in the same direction. This exclusion can be viewed as a "repulsive" mechanism for spins in the same direction; its effect is opposite that required to align the electrons responsible for the magnetization in the iron domains. However, other electrons with spins in the opposite direction, primarily in the fourth *s* atomic shell, interact at close range with the magnetization electrons, and this interaction is attractive. Because of the attractive effect of their opposite spins, these *s*-shell electrons influence the magnetization electrons of a number of the iron atoms and align them with each other.

A simple empirical representation of the effect of such exchange forces invokes the idea of an effective internal, or molecular, field $\boldsymbol{H}_{int}$, which is proportional in size to the magnetization $\boldsymbol{M}$; that is, $\boldsymbol{H}_{int} = \lambda \boldsymbol{M}$ in which λ is an empirical parameter. The resulting magnetization $\boldsymbol{M}$ equals $\chi_p(\boldsymbol{H} + \lambda \boldsymbol{M})$, in which χ_p is the susceptibility that the substance would have in the absence of the internal field. Assuming that $\chi_p = C/T$, corresponding to Curie's law, the equation $M = C(H + \lambda M)/T$ has the solution $\chi = M/H = C/(T - C\lambda) = C/(T - Tc)$. This result, the Curie–Weiss law, is valid at temperatures greater than the Curie temperature T_c (see below); at such temperatures the substance is still paramagnetic because the magnetization is zero when the field is zero. The internal field, however, makes the susceptibility larger than that given by the Curie law. A plot of $1/\chi$ against T still gives a straight line, as shown in Figure 46, but $1/\chi$ becomes zero when the temperature reaches the Curie temperature.

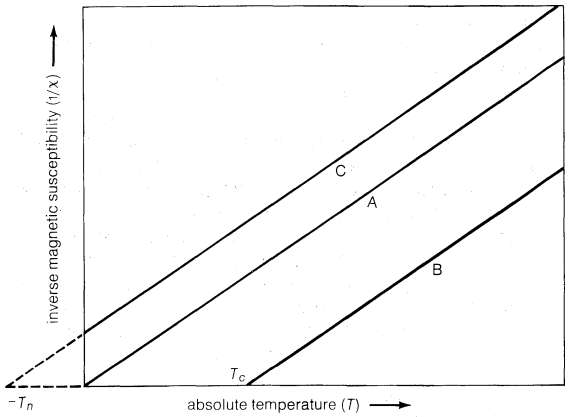


Figure 46: Plot of $1/\chi$. (A) Curie's law. (B) Curie–Weiss law for a ferromagnet with Curie temperature T_c. (C) Curie–Weiss law for an antiferromagnetic substance.

Since $1/\chi = H/M$, $\boldsymbol{M}$ at this temperature must be finite even when the magnetic field is zero. Thus, below the Curie temperature, the substance exhibits a spontaneous magnetization $\boldsymbol{M}$ in the absence of an external field, the essential property of a ferromagnet. Table 4 gives Curie temperature values for various ferromagnetic substances.

Curie temperature as the point of transition

Table 4: Curie Temperatures for Some Ferromagnetic Substances

Iron (Fe)	1,043 K
Cobalt (Co)	1,404 K
Nickel (Ni)	631 K
Gadolinium (Gd)	293 K
"Bismanol" (MnBi)	633 K
Manganese arsenide (MnAs)	318 K

In the ferromagnetic phase below the Curie temperature, the spontaneous alignment is still resisted by random thermal energy, and the spontaneous magnetization $\boldsymbol{M}$ is a function of temperature. The magnitude of $\boldsymbol{M}$ can be found from the paramagnetic equation for the reduced magnetization $M/M_s = f(mB/kT)$ by replacing B with $\mu(H + \lambda M)$. This gives an equation that can be solved numerically if the function f is known. When H equals zero, the curve of (M/M_s) should be a unique function of the ratio (T/T_c) for all substances that have the same function f. Such a curve is shown in Figure 47, together with experimental results for nickel and a nickel–copper alloy.

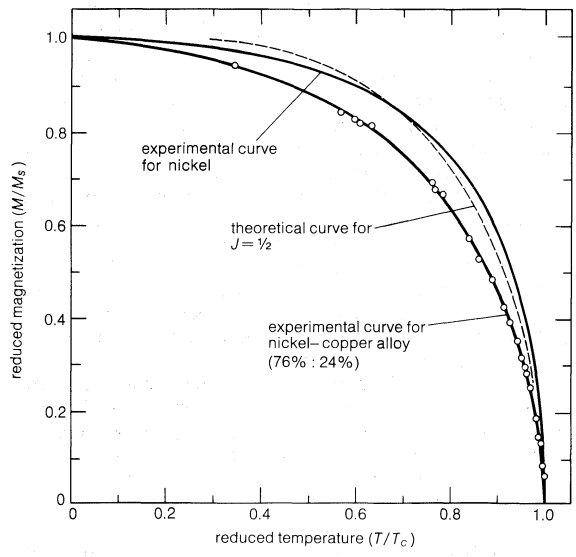


Figure 47: The reduced magnetization M/M_s as a function of reduced temperature T/T_c for a magnet.

The molecular field theory explains the existence of a ferromagnetic phase and the presence of spontaneous magnetization below the Curie temperature. The dependence of the magnetization on the external field is, however, more complex than the Curie–Weiss theory predicts. The magnetization curve is shown in Figure 48 for iron, with the field B in the iron plotted against the external field H. The variation is nonlinear, and B reaches its saturation value S in small fields. The relative permeability $B/\mu_0 H$ attains values of 10^3 to 10^4 in contrast to an ordinary paramagnet, for which μ is about 1.001 at room temperature. On reducing the external field H, the field B does not return along the magnetization curve. Even at $H = 0$, its value is not far below the saturation value.

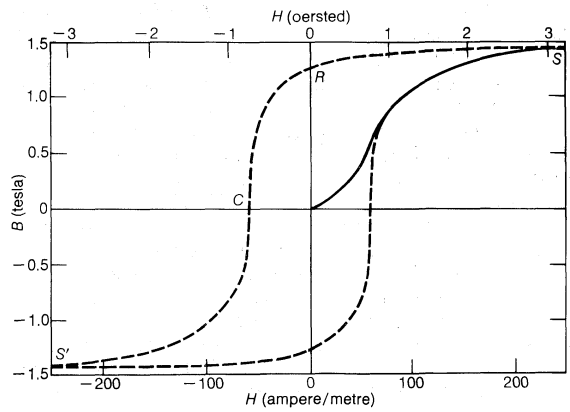


Figure 48: The magnetization curve (solid curve) and hysteresis loop (broken curve) for a ferromagnet.

When $H = 0$ (labeled R in the figure), the magnetic field constitutes what is termed the residual flux density, and the retention of magnetization in zero field is called remanence. When the external field is reversed, the value of B falls and passes through zero (point C) at a field strength known as the coercive force. Further increase in the reverse field H sets up a reverse field B that again quickly reaches a saturation value S'. Finally, as the reverse field is removed and a positive field applied, B traces out the lower broken line back to a positive saturation value. Further cycles of H retrace the broken curve, which is known as the hysteresis curve, because the change in B always lags behind the change in H. The hysteresis curve is not unique unless saturation is attained in each direction; interruption and reversal of the cycle at an intermediate field strength results in a hysteresis curve of smaller size.

Remanence

To explain ferromagnetic phenomena, Weiss suggested that a ferromagnetic substance contains many small regions (called domains), in each of which the substance is magnetized locally to saturations in some direction. In the unmagnetized state, such directions are distributed at random or in such a way that the net magnetization of the whole sample is zero. Application of an external field changes the direction of magnetization of part or all of the domains, setting up a net magnetization parallel to the field. In a paramagnetic substance, atomic dipoles are oriented on a microscopic scale. In contrast, the magnetization of a ferromagnetic substance involves the reorientation of the magnetization of the domains on a macroscopic scale; large changes occur in the net magnetization even when very small fields are applied. Such macroscopic changes are not immediately reversed when the size of the field is reduced or when its direction is changed. This accounts for the presence of hysteresis and for the finite remanent magnetization.

The technological applications of ferromagnetic substances are extensive, and the size and shape of the hysteresis curve are of great importance. A good permanent magnet must have a large spontaneous magnetization in zero field (*i.e.*, a high retentivity) and a high coercive force to prevent its being easily demagnetized by an external field. Both of these imply a "fat," almost rectangular hysteresis loop, typical of a hard magnetic material. On the other hand, ferromagnetic substances subjected to alternating fields, as in a transformer, must have a "thin" hysteresis loop because of an energy loss per cycle that is determined by the area enclosed by the hysteresis loop. Such substances are easily magnetized and demagnetized and are known as soft magnetic materials.

ANTIFERROMAGNETISM

Differences between ferromagnetic and antiferromagnetic substances

In substances known as antiferromagnets, the mutual forces between pairs of adjacent atomic dipoles are caused by exchange interactions, but the forces between adjacent atomic dipoles have signs opposite those in ferromagnets. As a result, adjacent dipoles tend to line up antiparallel to each other instead of parallel. At high temperatures the material is paramagnetic, but below a certain characteristic temperature the dipoles are aligned in an ordered and antiparallel manner. The transition temperature T_n is known as the Néel temperature, after the French physicist Louis-Eugène-Félix Néel, who proposed this explanation of the magnetic behaviour of such materials in 1936. Values of the Néel temperature for some typical antiferromagnetic substances are given in Table 5.

Table 5: Néel Temperature of Antiferromagnetic Substances

Chromium (Cr)	311 K
Manganese fluoride (MnF_2)	67 K
Nickel fluoride (NiF_2)	73 K
Manganese oxide (MnO)	116 K
Ferrous oxide (FeO)	198 K

The ordered antiferromagnetic state is naturally more complicated than the ordered ferromagnetic state, since there must be at least two sets of dipoles pointing in opposite directions. With an equal number of dipoles of the same size on each set, there is no net spontaneous magnetization on a macroscopic scale. For this reason, antiferromagnetic substances have few commercial applications. In most insulating chemical compounds, the exchange forces between the magnetic ions are of an antiferromagnetic nature.

FERRIMAGNETISM

Lodestone, or magnetite (Fe_3O_4), belongs to a class of substances known as ferrites. Ferrites and some other classes of magnetic substances discovered more recently possess many of the properties of ferromagnetic materials, including spontaneous magnetization and remanence. Unlike the ferromagnetic metals, they have low electric conductivity, however. In alternating magnetic fields, this greatly reduces the energy loss resulting from eddy currents. Since these losses rise with the frequency of the alternating field, such substances are of much importance in the electronics industry.

Low electric conductivity of ferrites

A notable property of ferrites and associated materials is that the bulk spontaneous magnetization, even at complete magnetic saturation, does not correspond to the value expected if all the atomic dipoles are aligned parallel to each other. The explanation was put forward in 1948 by Néel, who suggested that the exchange forces responsible for the spontaneous magnetization were basically antiferromagnetic in nature and that in the ordered state they contained two (or more) sublattices spontaneously magnetized in opposite directions. In contrast to the simple antiferromagnetic substances considered above, however, the sizes of the magnetization on the two sublattices are unequal, giving a resultant net magnetization parallel to that of the sublattice with the larger moment. For this phenomenon Néel coined the name ferrimagnetism, and substances that exhibit it are called ferrimagnetic materials.

(B.Ble./E.Ka./S.McG.)

Historical survey

Electric and magnetic forces have been known since antiquity, but they were regarded as separate phenomena for centuries. Magnetism was studied experimentally at least as early as the 13th century; the properties of the magnetic compass undoubtedly aroused interest in the phenomenon. Systematic investigations of electricity were delayed until the invention of practical devices for producing electric charge and currents. As soon as inexpensive, easy-to-use sources of electricity became available, scientists produced a wealth of experimental data and theoretical insights. As technology advanced, they studied, in turn, magnetism and electrostatics, electric currents and conduction, electrochemistry, magnetic and electric induction, the interrelationship between electricity and magnetism, and finally the fundamental nature of electric charge.

EARLY OBSERVATIONS AND APPLICATIONS OF ELECTRIC AND MAGNETIC PHENOMENA

The ancient Greeks knew about the attractive force of both magnetite and rubbed amber. Magnetite, a magnetic

oxide of iron mentioned in Greek texts as early as 800 BC, was mined in the province of Magnesia in Thessaly. Thales of Miletus, who lived nearby, may have been the first Greek to study magnetic forces. He apparently knew that magnetite attracts iron and that rubbing amber (a fossil tree resin that the Greeks called *ēlektron*) would make it attract such lightweight objects as feathers. According to Lucretius, the Roman author of the philosophical poem *De rerum natura* ("On the Nature of Things") in the 1st century BC, the term magnet was derived from the province of Magnesia. Pliny the Elder, however, attributes it to the supposed discoverer of the mineral, the shepherd Magnes, "the nails of whose shoes and the tip of whose staff stuck fast in a magnetic field while he pastured his flocks."

The magnetic compass

The oldest practical application of magnetism was the magnetic compass, but its origin remains unknown. Some historians believe it was used in China as far back as the 26th century BC; others contend that it was invented by the Italians or Arabs and introduced to the Chinese during the 13th century AD. The earliest extant European reference is by Alexander Neckam (*d.* 1217) of England.

The first experiments with magnetism are attributed to Petrus Peregrinus de Maricourt, a French crusader and engineer. In his oft-cited *Epistola de magnete* (1269; "Letter on the Magnet"), Peregrinus describes having placed a thin iron rectangle on different parts of a spherically shaped piece of magnetite (or lodestone) and marked the lines along which it set itself. The lines formed a set of meridians of longitude passing through two points at opposite ends of the stone, in much the same way as the lines of longitude on the Earth's surface intersect at the North and South poles. By analogy, Peregrinus called the points the poles of the magnet. He further noted that, when a magnet is cut into pieces, each piece still has two poles. He also observed that unlike poles attract each other and that a strong magnet can reverse the polarity of a weaker one.

EMERGENCE OF THE MODERN SCIENCES OF ELECTRICITY AND MAGNETISM

The founder of the modern sciences of electricity and magnetism was William Gilbert, physician to both Elizabeth I and James I of England. Gilbert spent 17 years experimenting with magnetism and, to a lesser extent, electricity. He assembled the results of his experiments and all of the available knowledge on magnetism in the treatise *De Magnete, Magneticisque Corporibus, et de Magno Magnete Tellure* ("On the Magnet and Magnetic Bodies, and on That Great Magnet the Earth"). As suggested by the title, Gilbert described the Earth as a huge magnet. He introduced the term electric for the force between two objects charged by friction and showed that frictional electricity occurs in many common materials. He also noted one of the primary distinctions between magnetism and electricity: the force between magnetic objects tends to align the objects relative to each other and is affected only slightly by most intervening objects, while the force between electrified objects is primarily a force of attraction or repulsion between the objects and is grossly affected by intervening matter. Gilbert attributed the electrification of a body by friction to the removal of a fluid, or "humour," which then left an "effluvium," or atmosphere, around the body. The language is quaint, but, if the "humour" is renamed "charge" and the "effluvium" renamed "electric field," Gilbert's notions closely approach modern ideas.

Pioneering efforts. During the 17th and early 18th centuries, as better sources of charge were developed, the study of electric effects became increasingly popular. The first machine to generate an electric spark was built in 1663 by Otto von Guericke, a German physicist and engineer. Guericke's electric generator consisted of a sulfur globe mounted on an iron shaft. The globe could be turned with one hand and rubbed with the other. Electrified by friction, the sphere alternately attracted and repulsed light objects from the floor.

Stephen Gray, a British chemist, is credited with discovering that electricity can flow (1729). He found that corks stuck in the ends of glass tubes become electrified when the tubes are rubbed. He also transmitted electricity approximately 150 metres through a hemp thread supported by silk cords and, in another demonstration, sent electricity even farther through metal wire. Gray concluded that electricity flowed everywhere.

From the mid-18th through the early 19th centuries, scientists believed that electricity was composed of fluid. In 1733 Charles François de Cisternay DuFay, a French chemist, announced that electricity consisted of two fluids: "vitreous" (from the Latin for "glass"), or positive, electricity; and "resinous," or negative, electricity. When DuFay electrified a glass rod, it attracted nearby bits of cork. Yet, if the rod touched the pieces of cork, the cork fragments were repelled and also repelled one another. DuFay accounted for this phenomenon by explaining that, in general, matter was neutral because it contained equal quantities of both fluids; if, however, friction separated the fluids in a substance and left it imbalanced, the substance would attract or repel other matter.

Invention of the Leyden jar. In 1745 a cheap and convenient source of electric sparks was invented by Pieter van Musschenbroek, a physicist and mathematician in Leiden, Neth. Later called the Leyden jar, it was the first device that could store large amounts of electric charge. (E. Georg von Kleist, a German cleric, independently developed the idea for such a device, but did not investigate it as thoroughly as did Musschenbroek.) The Leyden jar devised by the latter consisted of a glass vial that was partially filled with water and contained a thick conducting wire capable of storing a substantial amount of charge. One end of this wire protruded through the cork that sealed the opening of the vial. The Leyden jar was charged by bringing this exposed end of the conducting wire into contact with a friction device that generated static electricity.

Within a year after the appearance of Musschenbroek's device, William Watson, an English physician and scientist, constructed a more sophisticated version of the Leyden jar; he coated the inside and outside of the container with metal foil to improve its capacity to store charge. Watson transmitted an electric spark from his device through a wire strung across the River Thames at Westminster Bridge in 1747.

The Leyden jar revolutionized the study of electrostatics. Soon "electricians" were earning their living all over Europe demonstrating electricity with Leyden jars. Typically, they killed birds and animals with electric shock or sent charges through wires over rivers and lakes. In 1746 the abbé Jean-Antoine Nollet, a physicist who popularized science in France, discharged a Leyden jar in front of King Louis XV by sending current through a chain of 180 Royal Guards. In another demonstration, Nollet used wire made of iron to connect a row of Carthusian monks more than a kilometre long; when a Leyden jar was discharged, the white-robed monks reportedly leapt simultaneously into the air.

In the United States, Benjamin Franklin sold his printing house, newspaper, and almanac to spend his time conducting electricity experiments. In 1752 Franklin proved that lightning was an example of electric conduction by flying a silk kite during a thunderstorm. He collected electric charge from a cloud by means of wet twine attached to a key and thence to a Leyden jar. He then used the accumulated charge from the lightning to perform electric experiments. Franklin enunciated the law now known as the conservation of charge (the net sum of the charges within an isolated region is always constant). Like Watson, he disagreed with DuFay's two-fluid theory. Franklin argued that electricity consisted of two states of one fluid, which is present in everything. A substance containing an unusually large amount of the fluid would be "plus," or positively charged. Matter with less than a normal amount of fluid would be "minus," or negatively charged.

Franklin's one-fluid theory

Franklin's one-fluid theory, which dominated the study of electricity for 100 years, is essentially correct because most currents are the result of moving electrons. At the same time, however, fundamental particles have both negative and positive charges and, in this sense, DuFay's two-fluid picture is correct.

Joseph Priestley, an English physicist, summarized all available data on electricity in his book *History and Present*

State of Electricity (1767). He repeated one of Franklin's experiments, in which the latter had dropped small corks into a highly electrified metal container and found that they were neither attracted nor repelled. The lack of any charge on the inside of the container caused Priestley to recall Newton's law that there is no gravitational force on the inside of a hollow sphere. From this, Priestley inferred that the law of force between electric charges must be the same as the law for gravitational force—*i.e.,* that the force between masses diminishes with the inverse square of the distance between the masses. Although they were expressed in qualitative and descriptive terms, Priestley's laws are still valid today. Their mathematics was clarified and developed extensively between 1767 and the mid-19th century as electricity and magnetism became precise, quantitative sciences.

Formulation of the quantitative laws of electrostatics and magnetostatics. Charles-Augustin de Coulomb established electricity as a mathematical science during the latter half of the 18th century. He transformed Priestley's descriptive observations into the basic quantitative laws of electrostatics and magnetostatics. He also developed the mathematical theory of electric force and invented the torsion balance that was to be used in electricity experiments for the next 100 years. Coulomb used the balance to measure the force between magnetic poles and between electric charges at varying distances. In 1785 he announced his quantitative proof that electric and magnetic forces vary, like gravitation, inversely as the square of the distance (see above *General considerations*). Thus, according to Coulomb's law, if the distance between two charged masses is doubled, the electric force between them is reduced to a fourth. (The English physicist Henry Cavendish, as well as John Robison of Scotland, had made quantitative determinations of this principle before Coulomb, but they had not published their work.)

The mathematicians Siméon-Denis Poisson of France and Carl Friedrich Gauss of Germany extended Coulomb's work during the 18th and early 19th centuries. Poisson's equation (published in 1813) and the law of charge conservation contain in two lines virtually all the laws of electrostatics. The theory of magnetostatics, which is the study of steady-state magnetic fields, also was developed from Coulomb's law. Magnetostatics uses the concept of a magnetic potential analogous to the electric potential (*i.e.,* magnetic poles are postulated with properties analogous to electric charges).

Faraday's verification of the inverse square law

Michael Faraday built upon Priestley's work and conducted an experiment that verified quite accurately the inverse square law. Faraday's experiment involving the use of a metal ice pail and a gold-leaf electroscope was the first precise quantitative experiment on electric charge. In Faraday's time, the gold-leaf electroscope was used to indicate the electric state of a body. This type of apparatus consists of two thin leaves of gold hanging from an insulated metal rod that is mounted inside a metal box. When the rod is charged, the leaves repel each other and the deflection indicates the size of the charge. Faraday began his experiment by charging a metal ball suspended on an insulating silk thread. He then connected the gold-leaf electroscope to a metal ice pail resting on an insulating block and lowered the charged ball into the pail. The electroscope reading increased as the ball was lowered into the pail and reached a steady value once the ball was within the pail. When the ball was withdrawn without touching the pail, the electroscope reading fell to zero. Yet, when the ball touched the bottom of the pail, the reading remained at its steady value. On removal, the ball was found to be completely discharged. Faraday concluded that the electric charge produced on the outside of the pail, when the ball was inside but not in contact with it, was exactly equal to the initial charge on the ball. He then inserted into the pail other objects, such as a set of concentric pails separated from one another with various insulating materials like sulfur. In each case, the electroscope reading was the same once the ball was completely within the pail. From this, Faraday concluded that the total charge of the system was an invariable quantity equal

By courtesy of *Scientific American*, December 1988

Figure 49: *Widespread use of electricity in the 19th century.*
The study and use of electricity affected nearly every area of 19th-century life, even grand opera. In 1888 two singers in a performance of Charles Gounod's *Faust* staged a spectacular duel scene by forming an electric circuit. Their swords were connected to the poles of a battery under the stage floor via wires hidden beneath their costumes, copper nails in their shoes, and metal plates on the stage floor. Each time the swords touched, they sparkled and crackled like lightning.

to the initial charge of the ball. The present-day belief that conservation is a fundamental property of charge rests not only on the experiments of Franklin and Faraday but also on its complete agreement with all observations in electric engineering, quantum electrodynamics, and experimental electricity. With Faraday's work, the theory of electrostatics was complete.

FOUNDATIONS OF ELECTROCHEMISTRY AND ELECTRODYNAMICS

Development of the battery. The invention of the battery in 1800 made possible for the first time major advances in the theories of electric current and electrochemistry. Both science and technology developed rapidly as a direct result, leading some to call the 19th century the age of electricity.

The development of the battery was the accidental result of biological experiments conducted by Luigi Galvani. Galvani, a professor of anatomy at the Bologna Academy of Science, was interested in electricity in fish and other animals. One day he noticed that electric sparks from an electrostatic machine caused muscular contractions in a dissected frog that lay nearby. At first, Galvani assumed that the phenomenon was the result of atmospheric electricity because similar effects could be observed during lightning storms. Later, he discovered that whenever a piece of metal connected the muscle and nerve of the frog, the muscle contracted. Although Galvani realized that some metals appeared to be more effective than others in producing this effect, he concluded incorrectly that the metal was transporting a fluid, which he identified with animal electricity, from the nerve to the muscle. Galvani's observations, published in 1791, aroused considerable controversy and speculation.

Alessandro Volta, a physicist at the nearby University of Pavia, had been studying how electricity stimulates the senses of touch, taste, and sight. When Volta put a metal coin on top of his tongue and another coin of a different metal under his tongue and connected their surfaces with a wire, the coins tasted salty. Like Galvani, Volta assumed that he was working with animal electricity until 1796 when he discovered that he could also produce a current when he substituted a piece of cardboard soaked in brine for his tongue. Volta correctly conjectured that the effect was caused by the contact between metal and a moist body. Around 1800 he constructed what is now known as a voltaic pile consisting of layers of silver, moist cardboard, and zinc, repeated in that order, beginning and ending with a different metal. When he joined the silver and the zinc with a wire, electricity flowed continuously through the wire. Volta confirmed that the effects of his pile were equivalent in every way to those of static electricity. Within 20 years, galvanism, as electricity produced by a chemical reaction was then called, became unequivocally linked to static electricity. More important, Volta's invention provided the first source of continuous electric current. This rudimentary form of battery produced a smaller voltage than the Leyden jar, but it was easier to use because it could supply a steady current and did not have to be recharged.

Voltaic pile

The controversy between Galvani, who mistakenly thought that electricity originated in the animal's nerve, and Volta, who realized that it came from the metal, divided scientists into two camps. Galvani was supported by Alexander von Humboldt in Germany, while Volta was backed by Coulomb and other French physicists.

Within six weeks of Volta's report, two English scientists, William Nicholson and Anthony Carlisle, used a chemical battery to discover electrolysis (the process in which an electric current produces a chemical reaction) and initiate the science of electrochemistry. In their experiment the two employed a voltaic pile to liberate hydrogen and oxygen from water. They attached each end of the pile to brass wires and placed the opposite ends of the wires into salt water. The salt made the water a conductor. Hydrogen gas accumulated at the end of one wire; the end of the other wire was oxidized. Nicholson and Carlisle discovered that the amount of hydrogen and oxygen set free by the current was proportional to the amount of current used.

Discovery of electrolysis

By 1809 the English chemist Humphry Davy had used a stronger battery to free for the first time several very active metals—sodium, potassium, calcium, strontium, barium, and magnesium—from their liquid compounds. Faraday, who was Davy's assistant at the time, studied electrolysis quantitatively and showed that the amount of energy needed to separate a gram of a substance from its compound is closely related to the atomic weight of the substance. Electrolysis became a method of measuring electric current; and the quantity of charge that releases a gram atomic weight of a simple element is now called a faraday in his honour.

Once scientists were able to produce currents with a battery, they could study the flow of electricity quantitatively. Because of the battery, the German physicist Georg Simon Ohm was able experimentally in 1827 to quantify precisely a problem that Cavendish could only investigate qualitatively some 50 years earlier—namely, the ability of a material to conduct electricity. The result of this work—Ohm's law—explains how the resistance to the flow of charge depends on the type of conductor and on its length and diameter. According to Ohm's formulation, the current flow through a conductor is directly proportional to the potential difference, or voltage, and inversely proportional to the resistance—that is, $i = V/R$. Thus, doubling the length of an electric wire doubles its resistance, while doubling the cross-sectional area of the wire reduces the resistance by a half. Ohm's law is probably the most widely used equation in electric design.

Experimental and theoretical studies of electromagnetic phenomena. One of the great turning points in the development of the physical sciences was Hans Christian Ørsted's announcement in 1820 that electric currents produce magnetic effects. (Ørsted made his discovery while lecturing to a class of physics students. He placed by chance a wire carrying current near a compass needle and was surprised to see the needle swing at right angles to the wire.) Ørsted's fortuitous discovery proved that electricity and magnetism are linked. His finding, together with Faraday's subsequent discovery that a changing magnetic field produces an electric current in a nearby circuit, formed the basis of both James Clerk Maxwell's unified theory of electromagnetism and most of modern electrotechnology.

Verification of the link between electricity and magnetism

Once Ørsted's experiment had revealed that electric currents have magnetic effects, scientists realized that there must be magnetic forces between the currents. They began studying the forces immediately. A French physicist, François Arago, observed in 1820 that an electric current will orient unmagnetized iron filings in a circle around the wire. That same year, another French physicist, André-Marie Ampère, developed Ørsted's observations in quantitative terms. Ampère showed that two parallel wires carrying electric currents attract and repel each other like magnets. If the currents flow in the same direction, the wires attract each other; if they flow in opposite directions, the wires repel each other. From this experiment, Ampère was able to express the right-hand rule for the direction of the force on a current in a magnetic field. He also established experimentally and quantitatively the laws of magnetic force between electric currents. He suggested that internal electric currents are responsible for permanent magnets and for highly magnetizable materials like iron. With Arago, he demonstrated that steel needles become more strongly magnetic inside a coil carrying an electric current. Experiments on small coils showed that, at large distances, the forces between two such coils are similar to those between two small bar magnets and, moreover, that one coil can be replaced by a bar magnet of suitable size without changing the forces. The magnetic moment of this equivalent magnet was determined by the dimensions of the coil, its number of turns, and the current flowing around it.

William Sturgeon of England and Joseph Henry of the United States used Ørsted's discovery to develop electromagnets during the 1820s. Sturgeon wrapped 18 turns of bare copper wire around a U-shaped iron bar. When he turned on the current, the bar became an electromagnet capable of lifting 20 times its weight. When the current was turned off, the bar was no longer magnetized. Henry

repeated Sturgeon's work in 1829, using insulated wire to prevent short-circuiting. Using hundreds of turns, Henry created an electromagnet that could lift more than one ton of iron.

Ørsted's experiment showing that electricity could produce magnetic effects raised the opposite question as well: Could magnetism induce an electric current in another circuit? The French physicist Augustin-Jean Fresnel argued that since a steel bar inside a metallic helix can be magnetized by passing a current through the helix, the bar magnet in turn should create a current in an enveloping helix. In the following decade many ingenious experiments were devised, but the expectation that a steady current would be induced in a coil near the magnet resulted in experimenters either accidentally missing or not appreciating any transient electric effects caused by the magnet.

Faraday's discovery of electric induction

Faraday, the greatest experimentalist in electricity and magnetism of the 19th century and one of the greatest experimental physicists of all time, worked on and off for 10 years trying to prove that a magnet could induce electricity. In 1831 he finally succeeded by using two coils of wire wound around opposite sides of a ring of soft iron (Figure 50). The first coil was attached to a battery; when a current passed through the coil, the iron ring became magnetized. A wire from the second coil was extended to a compass needle a metre away, far enough so that it was not affected directly by any current in the first circuit. When the first circuit was turned on, Faraday observed a momentary deflection of the compass needle and its immediate return to its original position. When the primary current was switched off, a similar deflection of the compass needle occurred but in the opposite direction. Building on this observation in other experiments, Faraday showed that changes in the magnetic field around the first coil are responsible for inducing the current in the second coil. He also demonstrated that an electric current can be induced by moving a magnet, by turning an electromagnet on and off, and even by moving an electric wire in the Earth's magnetic field. Within a few months, Faraday built the first, albeit primitive, electric generator.

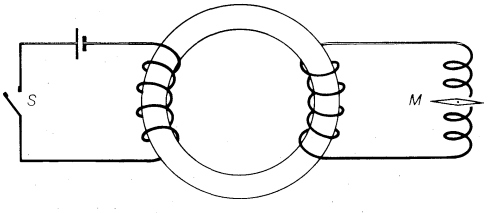


Figure 50: Faraday's magnetic induction experiment. When the switch S is closed in the primary circuit, a momentary current flows in the secondary circuit, giving a transient deflection of the compass needle M.

Henry had discovered electric induction quite independently in 1830, but his results were not published until after he had received news of Faraday's 1831 work, nor did he develop the discovery as fully as Faraday. In his paper of July 1832, Henry reported and correctly interpreted self-induction. He had produced large electric arcs from a long helical conductor when it was disconnected from a battery. When he had opened the circuit, the rapid decrease in the current had caused a large voltage between the battery terminal and the wire. As the wire lead was pulled away from the battery, the current continued to flow for a short time in the form of a bright arc between the battery terminal and the wire.

Faraday's thinking was permeated by the concept of electric and magnetic lines of force. He visualized that magnets, electric charges, and electric currents produce lines of force. When he placed a thin card covered with iron filings on a magnet, he could see the filings form chains from one end of the magnet to the other. He believed that these lines showed the directions of the forces and that electric current would have the same lines of force. The tension they build explains the attraction and repulsion of magnets and electric charges. Faraday had visualized magnetic curves as early as 1831 while working on his induction experiments; he wrote in his notes, "By magnetic curves I mean lines of magnetic forces which would be depicted by iron filings." Faraday opposed the prevailing idea that induction occurred "at a distance"; instead, he held that induction occurs along curved lines of force because of the action of contiguous particles. Later, he explained that electricity and magnetism are transmitted through a medium that is the site of electric or magnetic "fields," which make all substances magnetic to some extent.

Faraday was not the only researcher laying the groundwork for a synthesis between electricity, magnetism, and other areas of physics. On the continent of Europe, primarily in Germany, scientists were making mathematical connections between electricity, magnetism, and optics. The work of the physicists Franz Ernst Neumann, Wilhelm Eduard Weber, and H.F.E. Lenz belongs to this period. At the same time, Helmholtz and the English physicists William Thomson (later Lord Kelvin) and James Prescott Joule were clarifying the relationship between electricity and other forms of energy. Joule investigated the quantitative relationship between electric currents and heat during the 1840s and formulated the theory of the heating effects that accompany the flow of electricity in conductors. Helmholtz, Thomson, Henry, Gustav Kirchhoff, and Sir George Gabriel Stokes also extended the theory of the conduction and propagation of electric effects in conductors. In 1856 Weber and his German colleague, Rudolf Kohlrausch, determined the ratio of electric and magnetic units and found that it has the same dimensions as light and that it is almost exactly equal to its velocity. In 1857 Kirchhoff used this finding to demonstrate that electric disturbances propagate on a highly conductive wire with the speed of light.

Formulation of the classical theory of electromagnetism

The final steps in synthesizing electricity and magnetism into one coherent theory were made by Maxwell. He was deeply influenced by Faraday's work, having begun his study of the phenomena by translating Faraday's experimental findings into mathematics. (Faraday was self-taught and had never mastered mathematics.) In 1856 Maxwell developed the theory that the energy of the electromagnetic field is in the space around the conductors as well as in the conductors themselves. By 1864 he had formulated his own electromagnetic theory of light, predicting that both light and radio waves are electric and magnetic phenomena. While Faraday had discovered that changes in magnetic fields produce electric fields, Maxwell added the converse: changes in electric fields produce magnetic fields even in the absence of electric currents. Maxwell predicted that electromagnetic disturbances traveling through empty space have electric and magnetic fields at right angles to each other and that both fields are perpendicular to the direction of the wave. He concluded that the waves move at a uniform speed equal to the speed of light and that light is one form of electromagnetic wave. Their elegance notwithstanding, Maxwell's radical ideas were accepted by few outside England until 1886, when the German physicist Heinrich Hertz verified the existence of electromagnetic waves traveling at the speed of light; the waves he discovered are known now as radio waves.

Maxwell's four field equations (see above) represent the pinnacle of classical electromagnetic theory. Subsequent developments in the theory have been concerned either with the relationship between electromagnetism and the atomic structure of matter or with the practical and theoretical consequences of Maxwell's equations. His formulation has withstood the revolutions of relativity and quantum mechanics. His equations are appropriate for distances as small as 10^{-10} centimetres—100 times smaller than the size of an atom. The fusion of electromagnetic theory and quantum theory, known as quantum electrodynamics, is required only for smaller distances.

While the mainstream of theoretical activity concerning electric and magnetic phenomena during the 19th century was devoted to showing how they are interrelated, some scientists made use of them to discover new properties of materials and heat. Weber developed Ampère's suggestion that there are internal circulating currents of molecular size in metals. He explained how a substance loses its magnetic properties when the molecular magnets point in

random directions. Under the action of an external force, they may turn to point in the direction of the force; when all point in this direction, the maximum possible degree of magnetization is reached, a phenomenon known as magnetic saturation. In 1895 Pierre Curie of France discovered that a ferromagnetic substance has a specific temperature above which it ceases to be magnetic. Finally, superconductivity was discovered in 1900 by the German physicist Heike Kammerlingh-Onnes. In superconductivity electric conductors lose all resistance at very low temperatures.

Discovery of the electron and its ramifications. Although little of major importance was added to electromagnetic theory in the 19th century after Maxwell, the discovery of the electron in 1898 opened up an entirely new area of study: the nature of electric charge and of matter itself.

Studies of currents in vacuum tubes

The discovery of the electron grew out of studies of electric currents in vacuum tubes. Heinrich Geissler, a glassblower who assisted the German physicist Julius Plücker, improved the vacuum tube in 1854. Four years later, Plücker sealed two electrodes inside the tube, evacuated the air, and forced electric currents between the electrodes; he attributed the green glow that appeared on the wall of the tube to rays emanating from the cathode. From then until the end of the century, the properties of cathode-ray discharges were studied intensively. The work of the English physicist Sir William Crookes in 1879 indicated that the luminescence was a property of the electric current itself. Crookes concluded that the rays were composed of electrified charged particles. In 1898 another English physicist, Sir J.J. Thomson, identified a cathode ray as a stream of negatively charged particles, each having a mass $^1/_{1,836}$ smaller than that of a hydrogen ion. Thomson's discovery established the particulate nature of charge; his particles were later dubbed electrons.

Following the discovery of the electron, electromagnetic theory became an integral part of the theories of the atomic, subatomic, and subnuclear structure of matter. This shift in focus occurred as the result of an impasse between electromagnetic theory and statistical mechanics over attempts to understand radiation from hot bodies. Thermal radiation had been investigated in Germany by the physicist Wilhelm Wien between 1890 and 1900. Wien had virtually exhausted the resources of thermodynamics in dealing with this problem. Two British scientists, Lord Rayleigh (John William Strutt) and Sir James Hopwood Jeans, had by 1900 applied the newly developed science of statistical mechanics to the same problem. They obtained results that, though in agreement with Wien's thermodynamic conclusions (as distinct from his speculative extensions of thermodynamics), only partially agreed with experimental observations. The German physicist Max Planck attempted to combine the statistical approach with a thermodynamic approach. By concentrating on the necessity of fitting together the experimental data, he was led to the formulation of an empirical law that satisfied Wien's thermodynamic criteria and accommodated the experimental data. When Planck interpreted this law in terms of Rayleigh's statistical concepts, he concluded that radiation of frequency ν exists only in quanta of energy. Planck's result, including the introduction of the new universal constant $h\nu$ in 1900, marked the foundation of quantum mechanics and initiated a profound change in physical theory (see ATOMS: *Bohr's shell model*).

By 1900 it was apparent that Thomson's electrons were a universal constituent of matter and, thus, that matter is essentially electric in nature. As a result, in the early years of the 20th century, many physicists attempted to construct theories of the electromagnetic properties of metals, insulators, and magnetic materials in terms of electrons. In 1909 the Dutch physicist Hendrik Antoon Lorentz succeeded in doing so in *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat;* his work has since been modified by quantum theory.

Special theory of relativity. The other major conceptual advance in electromagnetic theory was the special theory of relativity. In Maxwell's time, a mechanistic view of the universe held sway. Sound was interpreted as an undulatory motion of the air, while light and other electromagnetic waves were regarded as undulatory motions of an intangible medium called ether. The question arose as to whether the velocity of light measured by an observer moving relative to ether would be affected by his motion. Albert Abraham Michelson and Edward W. Morley of the United States had demonstrated in 1887 that light in a vacuum on Earth travels at a constant speed which is independent of the direction of the light relative to the direction of the Earth's motion through the ether. Lorentz and Henri Poincaré, a French physicist, showed between 1900 and 1904 that the conclusions of Michelson and Morley were consistent with Maxwell's equations. On this basis, Lorentz and Poincaré developed a theory of relativity in which the absolute motion of a body relative to a hypothetical ether is no longer significant. Poincaré named the theory the principle of relativity in a lecture at the St. Louis Exposition in September 1904. Planck gave the first formulation of relativistic dynamics two years later. The most general formulation of the special theory of relativity, however, was put forth by Einstein in 1905, and the theory of relativity is usually associated with his name. Einstein postulated that the speed of light is a constant, independent of the motion of the source of the light, and showed how the Newtonian laws of mechanics would have to be modified. While Maxwell had synthesized electricity and magnetism into one theory, he had regarded them as essentially two interdependent phenomena; Einstein showed that they are two aspects of the same phenomenon.

Lorentz's and Poincaré's principle of relativity

Maxwell's equations, the special theory of relativity, the discovery of the electronic structure of matter, and the formulation of quantum mechanics all occurred before 1930. The quantum electrodynamics theory, developed between 1945 and 1955, subsequently resolved some minute discrepancies in the calculations of certain atomic properties. For example, the accuracy with which it is now possible to calculate one of the numbers describing the magnetic moment of the electron is comparable to measuring the distance between New York City and Los Angeles to within the thickness of a human hair. As a result, quantum electrodynamics is the most complete and precise theory of any physical phenomenon. The remarkable correspondence between theory and observation makes it unique among human endeavours.

DEVELOPMENT OF ELECTROMAGNETIC TECHNOLOGY

Electromagnetic technology began with Faraday's discovery of induction in 1831 (see above). His demonstration that a changing magnetic field induces an electric current in a nearby circuit showed that mechanical energy can be converted to electric energy. It provided the foundation for electric power generation, leading directly to the invention of the dynamo and the electric motor. Faraday's finding also proved crucial for lighting and heating systems.

The early electric industry was dominated by the problem of generating electricity on a large scale. Within a year of Faraday's discovery, a small hand-turned generator in which a magnet revolved around coils was demonstrated in Paris. In 1833 there appeared an English model that featured the modern arrangement of rotating the coils in the field of a fixed magnet. By 1850 generators were manufactured commercially in several countries. Permanent magnets were used to produce the magnetic field in generators until the principle of the self-excited generator was discovered in 1866. (A self-excited generator has stronger magnetic fields because it uses electromagnets powered by the generator itself.) In 1870 Zénobe Théophile Gramme, a Belgian manufacturer, built the first practical generator capable of producing a continuous current. It was soon found that the magnetic field is more effective if the coil windings are embedded in slots in the rotating iron armature. The slotted armature, still in use today, was invented in 1880 by the Swedish engineer Jonas Wenström. Faraday's 1831 discovery of the principle of the AC transformer was not put to practical use until the late 1880s when the heated debate over the merits of direct-current and alternating-current systems for power transmission was settled in favour of the latter.

Self-excited generator

At first, the only serious consideration for electric power was arc lighting, in which a brilliant light is emitted by

an electric spark between two electrodes. The arc lamp was too powerful for domestic use, however, and so it was limited to large installations like lighthouses, train stations, and department stores. Commercial development of an incandescent filament lamp, first invented in the 1840s, was delayed until a filament could be made that would heat to incandescence without melting and until a satisfactory vacuum tube could be built. The mercury pump, invented in 1865, provided an adequate vacuum, and a satisfactory carbon filament was developed independently by the English physicist Sir Joseph Wilson Swan and the American inventor Thomas A. Edison during the late 1870s. By 1880 both had applied for patents for their incandescent lamps, and the ensuing litigation between the two men was resolved by the formation of a joint company in 1883. Thanks to the incandescent lamp, electric lighting became an accepted part of urban life by 1900. Since then, the tungsten filament lamp, introduced during the early 1900s, has become the principal form of electric lamp, though more efficient fluorescent gas discharge lamps have found widespread use as well.

Electricity took on a new importance with the development of the electric motor. This machine, which converts electric energy to mechanical energy, has become an integral component of a wide assortment of devices ranging from kitchen appliances and office equipment to industrial robots and rapid-transit vehicles. Although the principle of the electric motor was devised by Faraday in 1821, no commercially significant unit was produced until 1873. In fact, the first important AC motor, built by the Serbian-American inventor Nikola Tesla, was not demonstrated in the United States until 1888. Tesla began producing his motors in association with the Westinghouse Electric Company a few years after DC motors had been installed in trains in Germany and Ireland. By the end of the 19th century, the electric motor had taken a recognizably modern form. Subsequent improvements have rarely involved radically new ideas; however, the introduction of better designs and new bearing, armature, magnetic, and contact materials has resulted in the manufacture of smaller, cheaper, and more efficient and reliable motors.

The modern communications industry is among the most spectacular products of electricity. Telegraph systems using wires and simple electrochemical or electromechanical receivers proliferated in western Europe and the United States during the 1840s. An operable cable was installed under the English Channel in 1865, and a pair of transatlantic cables were successfully laid a year later. By 1872 almost all of the major cities of the world were linked by telegraph.

Alexander Graham Bell patented the first practical telephone in the United States in 1876, and the first public telephone services were operating within a few years. In 1895 the British physicist Sir Ernest Rutherford advanced Hertz's scientific investigations of radio waves and transmitted radio signals for more than one kilometre. Guglielmo Marconi, an Italian physicist and inventor, established wireless communications across the Atlantic employing radio waves of approximately 300- to 3,000-metre wavelength in 1901. Broadcast radio transmissions were established during the 1920s.

Transmission of radio signals

Telephone transmissions by radio waves, the electric recording and reproduction of sound, and television were made possible by the development of the triode tube. This three-electrode tube, invented by the American engineer Lee De Forest, permitted for the first time the amplification of electric signals. Known as the Audion, this device played a pivotal role in the early development of the electronics industry.

The first telephone transmission via radio signals was made from Arlington, Va., to the Eiffel Tower in Paris in 1915; and a commercial radio telephone service between New York City and London was begun in 1927. Besides such efforts, most of the major developmental work of this period was tied to the radio and phonograph entertainment industries and the sound film industry. Rapid progress was made toward transmitting moving pictures, especially in Great Britain; just before World War II, the British Broadcasting Corporation inaugurated the first public television service. Today, many regions of the electromagnetic spectrum are used for communications, including microwaves in the frequency range of approximately 7×10^9 hertz for satellite communication links and infrared light at a frequency of about 3×10^{14} hertz for optical fibre communications systems.

Until 1939 the electronics industry was almost exclusively concerned with communications and broadcast entertainment. Scientists and engineers in Britain, Germany, France, and the United States did initiate research on radar systems capable of aircraft detection and antiaircraft fire-control during the 1930s, however, and this marked the beginning of a new direction for electronics. During World War II and after, the electronics industry made strides paralleled only by those of the chemical industry. Television became commonplace; and a broad array of new devices and systems, most notably the electronic digital computer, emerged.

The electronic revolution of the last half of the 20th century has been made possible in large part by the invention of the transistor (1947) and such subsequent developments as the integrated circuit. (For detailed coverage of these and other major advances, see ELECTRONICS.) This miniaturization and integration of circuit elements has led to a remarkable diminution in the size and cost of electronic equipment and an equally impressive increase in its reliability. (F.N.H.R./E.Ka./S.McG.)

BIBLIOGRAPHY

Electricity and magnetism: P.C.W. DAVIES, *The Forces of Nature,* 2nd ed. (1986), is an interesting, readable account. DONALD M. TROTTER, JR., "Capacitors," *Scientific American,* 259(1):86–90B (July 1988), provides insight into capacitor functions and their role in technology. DAVID N. SCHRAMM and GARY STEIGMAN, "Particle Accelerators Test Cosmological Theory," *Scientific American,* 258(6):66–72 (June 1988), discusses the fundamental constituents of nature. EDWARD M. PURCELL, *Electricity and Magnetism,* 2nd ed. (1985), is superbly illustrated and treats key principles and phenomena with remarkable insight. Many examples and problems on electricity and magnetism, as well as elementary discussions of vectors and other aspects of physics, are found in DAVID HALLIDAY and ROBERT RESNICK, *Fundamentals of Physics,* 3rd ed. (1988). Useful physics textbooks with illustrations, examples, and problems include RICHARD WOLFSON and JAY M. PASACHOFF, *Physics* (1987); and FRANCIS W. SEARS, MARK W. ZEMANSKY, and HUGH D. YOUNG, *University Physics,* 7th ed. (1987).

Electromagnetism: RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* vol. 2, *The Electromagnetic Field* (1964, reprinted 1977), is highly recommended for its lucid discussion of fundamentals. JOHN R. REITZ, FREDERICK J. MILFORD, and ROBERT W. CHRISTY, *Foundations of Electromagnetic Theory,* 3rd ed. (1979), is a fine, compact, college-level text using vector calculus; while JOHN DAVID JACKSON, *Classical Electrodynamics,* 2nd ed. (1975), is written at the graduate level. E. DURAND, *Électrostatique,* 3 vol. (1964–66), and *Magnétostatique* (1968), exhaustively treat analytical methods and solutions of a variety of problems in electrostatics and magnetostatics, including dielectric and magnetic materials and conduction.

Electric and magnetic properties of matter: HARALD A. ENGE, *Introduction to Nuclear Physics* (1966), provides an overview. The magnetic properties of solids are discussed in CHARLES KITTEL, *Introduction to Solid State Physics,* 6th ed. (1986). ROBERT EISBERG and ROBERT RESNICK, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles,* 2nd ed. (1985), broadly treats quantum mechanical effects in various phenomena, including magnetic properties such as ferromagnetism and electric properties such as conduction in solids. Reference books include *Reference Data for Engineers: Radio, Electronics, Computer, and Communications,* 7th ed. (1985), with data and discussion about electric and magnetic properties of matter; and *CRC Handbook of Chemistry and Physics* (annual), an indispensable handbook.

History: J.L. HEILBRON, *Electricity in the 17th and 18th Centuries: A Study of Early Modern Physics* (1979), provides a readable survey of significant developments, as does EDMUND WHITTAKER, *A History of the Theories of Aether and Electricity,* rev. and enlarged ed., 2 vol. (1951–53, reprinted 1973). CHARLES SINGER and T.I. WILLIAMS (eds.), *A History of Technology,* 8 vol. (1954–84), begins in the prehistoric period and concludes around 1950. (E.Ka./ S.McG.)

Electromagnetic Radiation

In terms of classical theory, electromagnetic radiation is the flow of energy through space at the universal speed of light in the form of electric and magnetic fields that make up an electromagnetic wave. In such a wave, time-varying electric and magnetic fields are mutually linked with each other at right angles and perpendicular to the direction of motion. An electromagnetic wave is characterized by its intensity and the frequency ν of the time variation of the electric and magnetic fields.

In terms of the modern quantum theory, electromagnetic radiation is the flow of photons (also called light quanta) through space. Photons are packets of energy $h\nu$ that always move with the universal speed of light. The symbol h is Planck's constant, while the value of ν is the same as that of the frequency of the electromagnetic wave of classical theory. Photons having the same energy $h\nu$ are all alike, and their number density corresponds to the intensity of the radiation. Electromagnetic radiation exhibits a multitude of phenomena as it interacts with charged particles in atoms, molecules, and larger objects of matter. These phenomena as well as the ways in which electromagnetic radiation is created and observed, the manner in which such radiation occurs in nature, and its technological uses depend on its frequency ν. The spectrum of frequencies of electromagnetic radiation extends from very low values over the range of radio waves, television waves, and microwaves to visible light and beyond to the substantially higher values of ultraviolet light, X rays, and gamma rays.

The basic properties and behaviour of electromagnetic radiation are discussed in this article, as are its various forms, including their sources, distinguishing characteristics, and practical applications. The article also traces the development of both the classical and quantum theories of radiation.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 111, 112, 127, 128, and 131, and the *Index.*

The article is divided into the following sections:

General considerations

OCCURRENCE AND IMPORTANCE

Close to 0.01 percent of the mass/energy of the entire universe occurs in the form of electromagnetic radiation. All human life is immersed in it and modern communications technology and medical services are particularly dependent on one or another of its forms. In fact, all living things on Earth depend on the electromagnetic radiation received from the Sun and on the transformation of solar energy by photosynthesis into plant life or by biosynthesis into zooplankton, the basic step in the food chain in oceans. The eyes of many animals, including those of humans, are adapted to be sensitive to and hence to see the most abundant part of the Sun's electromagnetic radiation—namely, light, which comprises the visible portion of its wide range of frequencies. Green plants also have high sensitivity to the maximum intensity of solar electromagnetic radiation, which is absorbed by a substance called chlorophyll that is essential for plant growth via photosynthesis.

Importance to life

Practically all the fuels that modern society uses—gas, oil, and coal—are stored forms of energy received from the Sun as electromagnetic radiation millions of years ago. Only the energy from nuclear reactors does not originate from the Sun.

Everyday life is pervaded by man-made electromagnetic radiation: food is heated in microwave ovens, airplanes are guided by radar waves, television sets receive electromagnetic waves transmitted by broadcasting stations, and infrared waves from heaters provide warmth. Infrared waves also are given off and received by automatic self-focusing cameras that electronically measure and set the correct distance to the object to be photographed. As soon as the Sun sets, incandescent or fluorescent lights are turned on to provide artificial illumination, and cities glow brightly with the colourful fluorescent and neon lamps of advertisement signs. Familiar too is ultraviolet radiation, which the eyes cannot see but whose effect is felt as pain from sunburn. Ultraviolet light represents a kind of electromagnetic radiation that can be harmful to life. Such is also true of X rays, which are important in medicine as they allow physicians to observe the inner parts of the body but exposure to which should be kept to a minimum. Less familiar are gamma rays, which come from nuclear reactions and radioactive decay and are part of the harmful high-energy radiation of radioactive materials and nuclear weapons.

THE ELECTROMAGNETIC SPECTRUM

The brief account of familiar phenomena given above surveyed electromagnetic radiation from small frequencies ν (long wave radios) to exceedingly high values of ν (gamma rays). Going from the ν values of radio waves to those of visible light is like comparing the thickness of this page with the distance of the Earth from the Sun, which represents an increase by a factor of a million billion. Similarly, going from the ν values of visible light to the very much larger ones of gamma rays represents another increase in frequency by a factor of a million billion. This extremely large range of ν values, called the electromagnetic spec-

Vast range of frequencies

trum, is shown in Figure 1, together with the common names used for its various parts, or regions.

The number ν is shared by both the classical and the modern interpretation of electromagnetic radiation. In classical language, ν is the frequency of the temporal changes in an electromagnetic wave. The frequency of a wave is related to its speed c and wavelength λ in the following way. If 10 complete waves pass by in one second, one observes 10 wriggles, and one says that the frequency of such a wave is $\nu = 10$ cycles per second (10 hertz [Hz]). If the wavelength of the wave is, say, $\lambda = 3$ centimetres, then it is clear that a wave train 30 centimetres long has passed in that one second to produce the 10 wriggles that were observed. Thus, the speed of the wave is 30 centimetres per second, and one notes that in general the speed is $c = \lambda\nu$. The speed of electromagnetic radiation of all kinds is the same universal constant that is defined to be exactly $c = 299{,}792{,}458$ metres per second (186,282 miles per second). The wavelengths of the classical electromagnetic waves in free space calculated from $c = \lambda\nu$ are also shown on the spectrum in Figure 1, as is the energy $h\nu$ of modern-day photons. One commonly uses as the unit of energy electron volt (eV), which is the energy that can be given to an electron by a one-volt battery. It is clear that the range of wavelengths λ and of photon energies $h\nu$ are equally as large as the spectrum of ν values.

Because the wavelengths and energy quanta $h\nu$ of electromagnetic radiation of the various parts of the spectrum are so different in magnitude, the sources of the radiations, the interactions with matter, and the detectors employed are correspondingly different. This is why the same electromagnetic radiation is called by different names in various regions of the spectrum.

In spite of these obvious differences of scale, all forms of electromagnetic radiation obey certain general rules that are well understood and that allow one to calculate with very high precision their properties and interactions with charged particles in atoms, molecules, and large objects.

Electric and magnetic fields

Electromagnetic radiation is, classically speaking, a wave of electric and magnetic fields propagating at the speed of light c through empty space. In this wave the electric and magnetic fields change their magnitude and direction each second. This rate of change is the frequency ν measured in cycles per second—namely, in hertz. The electric and magnetic fields are always perpendicular to one another and at right angles to the direction of propagation, as shown in Figure 2. There is as much energy carried by the electric component of the wave as by the magnetic component, and the energy is proportional to the square of the field strength.

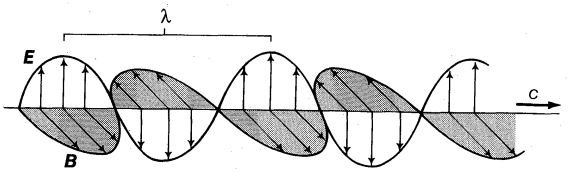


Figure 2: Radiation fields in which vectors ***E*** and ***B*** are perpendicular to each other and to the direction of propagation (see text).

GENERATION OF ELECTROMAGNETIC RADIATION

Electromagnetic radiation is produced whenever a charged particle, such as an electron, changes its velocity—*i.e.*, whenever it is accelerated or decelerated. The energy of the electromagnetic radiation thus produced comes from the charged particle and is therefore lost by it. A common example of this phenomenon is the oscillating charge or current in a radio antenna. The antenna of a radio transmitter is part of an electric resonance circuit in which the charge is made to oscillate at a desired frequency. An electromagnetic wave so generated can be received by a similar antenna connected to an oscillating electric circuit in the tuner that is tuned to that same frequency. The electromagnetic wave in turn produces an oscillating motion of charge in the receiving antenna. In general, one can say that any system which emits electromagnetic radiation of a given frequency can absorb radiation of the same frequency.

Such man-made transmitters and receivers become smaller with decreasing wavelength of the electromagnetic wave and prove impractical in the millimetre range. At even shorter wavelengths down to the wavelengths of X rays, which are one million times smaller, the oscillating charges arise from moving charges in molecules and atoms.

Systems or processes that generate electromagnetic radiation

One may classify the generation of electromagnetic radiation into two categories: (1) systems or processes that produce radiation covering a broad continuous spectrum of frequencies and (2) those that emit (and absorb) radiation of discrete frequencies that are characteristic of particular systems. The Sun with its continuous spectrum is an example of the first, while a radio transmitter tuned to one frequency exemplifies the second category.

Continuous spectra of electromagnetic radiation. Such spectra are emitted by any warm substance. Heat is the irregular motion of electrons, atoms, and molecules; the higher the temperature, the more rapid is the motion. Since electrons are much lighter than atoms, irregular thermal motion produces irregular oscillatory charge motion, which reflects a continuous spectrum of frequencies. Each oscillation at a particular frequency can be considered a tiny "antenna" that emits and receives electromagnetic radiation. As a piece of iron is heated to increasingly high temperatures, it first glows red, then yellow, and finally white. In short, all the colours of the visible spectrum are represented. Even before the iron begins to glow red, one can feel the emission of infrared waves by the heat sensation on the skin. A white-hot piece of iron also emits ultraviolet radiation, which can be detected by a photographic film.

Not all materials heated to the same temperature emit the same amount and spectral distribution of electromagnetic waves. For example, a piece of glass heated next to iron looks nearly colourless, but it feels hotter to the skin (it emits more infrared rays) than does the iron. This observation illustrates the rule of reciprocity: a body radiates strongly at those frequencies that it is able to absorb, because for both processes it needs the tiny antennas of that range of frequencies. Glass is transparent in the visible range of light because it lacks possible electronic absorption at these particular frequencies. As a consequence,

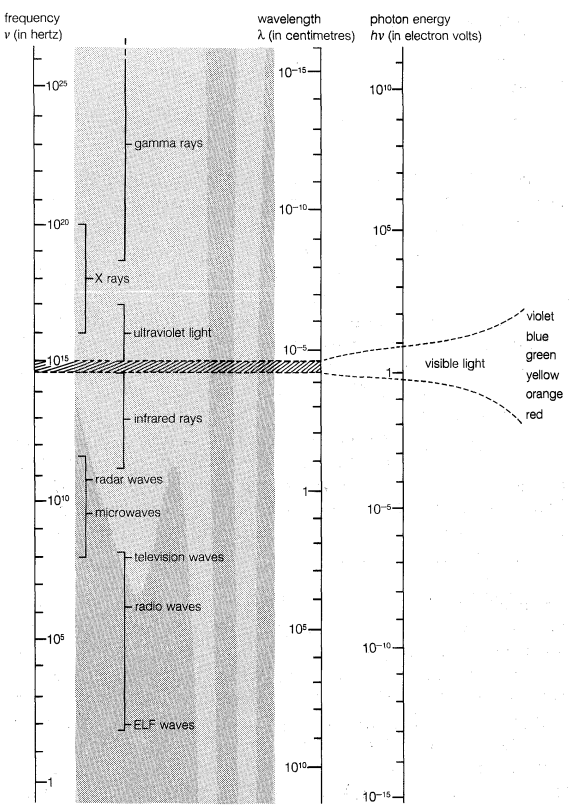


Figure 1: *Electromagnetic spectrum.*
The small visible range (shaded) is shown enlarged at the right.

glass cannot glow red because it cannot absorb red. On the other hand, glass is a better emitter/absorber in the infrared than iron or any other metal that strongly reflects such lower frequency electromagnetic waves. This selective emissivity and absorptivity is important for understanding the greenhouse effect (see below *The greenhouse effect of the atmosphere*) and many other phenomena in nature. The tungsten filament of a light bulb has a temperature of 2,500 K (4,040° F) and emits large amounts of visible light but relatively little infrared because metals, as mentioned above, have small emissivities in the infrared range. This is of course fortunate, since one wants light from a light bulb but not much heat. The light emitted by a candle originates from very hot carbon soot particles in the flame, which strongly absorb and thus emit visible light. By contrast, the gas flame of a kitchen range is pale, even though it is hotter than a candle flame, because of the absence of soot. Light from the stars originates from the high temperature of the gases at their surface. A wide spectrum of radiation is emitted from the Sun's surface, the temperature of which is about 6,000 K. The radiation output is 60 million watts for every square metre of solar surface, which is equivalent to the amount produced by an average-size commercial power-generating station that can supply electric power for about 30,000 households.

Blackbody radiation

The spectral composition of a heated body depends on the materials of which the body consists. That is not the case for an ideal radiator or absorber. Such an ideal object absorbs and thus emits radiation of all frequencies equally and fully. A radiator/absorber of this kind is called a blackbody, and its radiation spectrum is referred to as blackbody radiation, which depends on only one parameter, its temperature. Scientists devise and study such ideal objects because their properties can be known exactly. This information can then be used to determine and understand why real objects, such as a piece of iron or glass, a cloud, or a star, behave differently.

A good approximation of a blackbody is a piece of coal or, better yet, a cavity in a piece of coal that is visible through a small opening. There is one property of blackbody radiation which is familiar to everyone but which is actually quite mysterious. As the piece of coal is heated to higher and higher temperatures, one first observes a dull red glow, followed by a change in colour to bright red; as the temperature is increased further, the colour changes to yellow and finally to white. White is not itself a colour but rather the visual effect of the combination of all primary colours. The fact that white glow is observed at high temperatures means that the colour blue has been added to the ones observed at lower temperatures. This colour change with temperature is mysterious because one would expect, as the energy (or temperature) is increased, just more of the same and not something entirely different. For example, as one increases the power of a radio amplifier, one hears the music louder but not at a higher pitch.

Antennas of vibrating changes

The change in colour or frequency distribution of the electromagnetic radiation coming from heated bodies at different temperatures remained an enigma for centuries. The solution of this mystery by the German physicist Max Planck initiated the era of modern physics at the beginning of the 20th century. He explained the phenomenon by proposing that the tiny antennas in the heated body are quantized, meaning that they can emit electromagnetic radiation only in finite energy quanta of size hv. The universal constant h is called Planck's constant in his honour. For blue light $hv = 3$ eV, whereas $hv = 1.8$ eV for red light. Since high-frequency antennas of vibrating charges in solids have to emit larger energy quanta hv than lower-frequency antennas, they can only do so when the temperature, or the thermal atomic motion, becomes high enough. Hence, the average pitch, or peak frequency, of blackbody electromagnetic radiation increases with temperature.

The many tiny antennas in a heated chunk of material are, as noted above, to be identified with the accelerating and decelerating charges in the heat motion of the atoms of the material. There are other sources of continuous spectra of electromagnetic radiation that are not associated with heat but still come from accelerated or decelerated charges. X rays are, for example, produced by abruptly stopping rapidly moving electrons. This deceleration of the charges produces bremsstrahlung ("braking radiation"). In an X-ray tube, electrons moving with an energy of $E_{max} = 10{,}000$ to 50,000 eV (10–50 keV) are made to strike a piece of metal. The electromagnetic radiation produced by this sudden deceleration of electrons is a continuous spectrum extending up to the maximum photon energy $hv = E_{max}$.

By far the brightest continuum spectra of electromagnetic radiation come from synchrotron radiation sources. These are not well known because they are predominantly used for research and only recently have they been considered for commercial and medical applications. Because any change in motion is an acceleration, circulating currents of electrons produce electromagnetic radiation. When these circulating electrons move at relativistic speeds (*i.e.*, those approaching the speed of light), the brightness of the radiation increases enormously. This radiation was first observed at the General Electric Company in 1947 in an electron synchrotron (hence the name of this radiation), which is a type of particle accelerator that forces relativistic electrons into circular orbits using powerful magnetic fields. The intensity of synchrotron radiation is further increased more than a thousandfold by wigglers and undulators that move the beam of relativistic electrons to and fro by means of other magnetic fields.

The conditions for generating bremsstrahlung as well as synchrotron radiation exist in nature in various forms. Acceleration and capture of charged particles by the gravitational field of a star, black hole, or galaxy is a source of energetic cosmic X rays. Gamma rays are produced in other kinds of cosmic objects—namely, supernovae, neutron stars, and quasars.

Discrete-frequency sources and absorbers of electromagnetic radiation. These are commonly encountered in everyday life. Familiar examples of discrete-frequency electromagnetic radiation include the distinct colours of lamps filled with different fluorescent gases characteristic of advertisement signs, the colours of dyes and pigments, the bright yellow of sodium lamps, the blue-green hue of mercury lamps, and the specific colours of lasers.

Internal energy states

Sources of electromagnetic radiation of specific frequency are typically atoms or molecules. Every atom or molecule can have certain discrete internal energies, which are called quantum states. An atom or molecule can therefore change its internal energy only by discrete amounts. By going from a higher to a lower energy state, a quantum hv of electromagnetic radiation is emitted of a magnitude that is precisely the energy difference between the higher and lower state. Absorption of a quantum hv brings the atom from a lower to a higher state if hv matches the energy difference. All like atoms are identical, but all different chemical elements of the periodic table have their own specific set of possible internal energies. Therefore, by measuring the characteristic and discrete electromagnetic radiation that is either emitted or absorbed by atoms or molecules, one can identify which kind of atom or molecule is giving off or absorbing the radiation. This provides a means of determining the chemical composition of substances. Since one cannot subject a piece of a distant star to conventional chemical analysis, studying the emission or absorption of starlight is the only way to determine the composition of stars or of interstellar gases and dust.

The Sun, for example, not only emits the continuous spectrum of radiation that originates from its hot surface but also emits discrete radiation quanta hv that are characteristic of its atomic composition. Many of the elements can be detected at the solar surface, but the most abundant is helium. This is so because helium is the end product of the nuclear fusion reaction that is the fundamental energy source of the Sun. This particular element was named helium (from the Greek word *helios,* meaning "Sun") because its existence was first discovered by its characteristic absorption energies in the Sun's spectrum. The helium of the cooler outer parts of the solar atmosphere absorbs the characteristic light frequencies from the lower and hotter regions of the Sun.

The characteristic and discrete energies $h\nu$ found as emission and absorption of electromagnetic radiation by atoms and molecules extend to X-ray energies. As high-energy electrons strike the piece of metal in an X-ray tube, electrons are knocked out of the inner energy shell of the atoms. These vacancies are then filled by electrons from the second or third shell; emitted in the process are X rays having $h\nu$ values that correspond to the energy differences of the shells. One therefore observes not only the continuous spectrum of the bremsstrahlung discussed above but also X-ray emissions of discrete energies $h\nu$ that are characteristic of the specific elemental composition of the metal struck by the energetic electrons in the X-ray tube.

The discrete electromagnetic radiation energies $h\nu$ emitted or absorbed by all substances reflect the discreteness of the internal energies of all material things. This means that window glass and water are transparent to visible light; they cannot absorb these visible light quanta because their internal energies are such that no energy difference between a higher and a lower internal state matches the energy $h\nu$ of visible light. Figure 3 shows as an example the coefficient of absorption of water as a function of frequency ν of electromagnetic radiation. Above the scale of frequencies, the corresponding scales of photon energy $h\nu$ and wavelength λ are given. An absorption coefficient $\alpha = 10^{-4}$ cm^{-1} means that the intensity of electromagnetic radiation is only one-third its original value after passing through 100 metres of water. When $\alpha = 1$ cm^{-1}, only a layer one centimetre thick is needed to decrease the intensity to one-third its original value, and, for $\alpha = 10^3$ cm, a layer of water having a thickness of this page is sufficient to attenuate electromagnetic radiation by that much. The transparency of water to visible light, marked by the vertical dashed lines, is a remarkable feature that is significant for life on Earth.

Transparency of water to visible light

From J.D. Jackson, *Classical Electrodynamics*, copyright © 1975 by John Wiley & Sons, Inc.

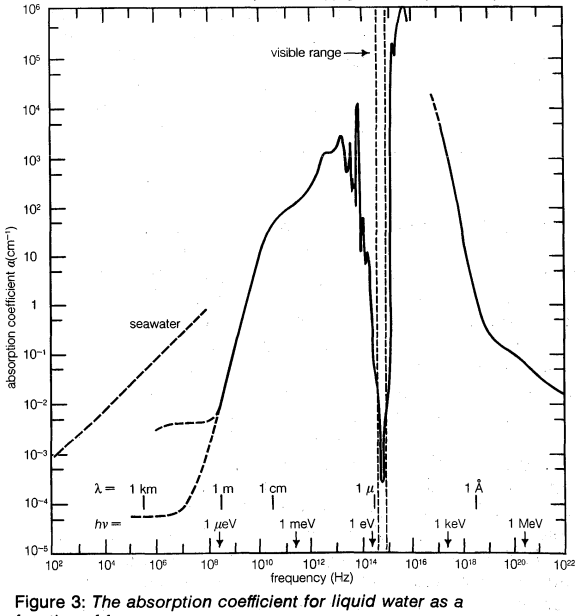


Figure 3: *The absorption coefficient for liquid water as a function of frequency.*
Also shown as abscissas are an energy scale (arrows) and a wavelength scale (vertical lines). The visible region of the frequency spectrum is indicated by the vertical dashed lines. The absorption coefficient for seawater is denoted by the dashed diagonal line at the left. The scales are logarithmic in both directions.

All things look so different and have different colours because of their different sets of internal discrete energies, which determine their interaction with electromagnetic radiation. The words *looking* and *colours* are associated with the human detectors of electromagnetic radiation, the eyes. Since there are instruments available for detecting electromagnetic radiation of any frequency, one can imagine that things "look" different at all energies of the spectrum because different materials have their own characteristic sets of discrete internal energies. Even the nuclei of atoms are composites of other elementary particles and thus can be excited to many discrete internal energy states. Since nuclear energies are much larger than atomic energies, the energy differences between internal energy states are substantially larger, and the corresponding electromagnetic radiation quanta $h\nu$ emitted or absorbed when nuclei change their energies are even bigger than those of X rays. Such quanta given off or absorbed by atomic nuclei are called gamma rays (see *The electromagnetic spectrum* above).

PROPERTIES AND BEHAVIOUR

Scattering, reflection, and refraction. If a charged particle interacts with an electromagnetic wave, it experiences a force proportional to the strength of the electric field and thus is forced to change its motion in accordance with the frequency of the electric field wave. In doing so, it becomes a source of electromagnetic radiation of the same frequency, as described in the previous section. The energy for the work done in accelerating the charged particle and emitting this secondary radiation comes from and is lost by the primary wave. This process is called scattering.

Since the energy density of the electromagnetic radiation is proportional to the square of the electric field strength and the field strength is caused by acceleration of a charge, the energy radiated by such a charge oscillator increases with the square of the acceleration. On the other hand, the acceleration of an oscillator depends on the frequency of the back-and-forth oscillation. The acceleration increases with the square of the frequency. This leads to the important result that the electromagnetic energy radiated by an oscillator increases very rapidly—namely, with the square of the square or, as one says, with the fourth power of the frequency. Doubling the frequency thus produces an increase in radiated energy by a factor of 16.

This rapid increase in scattering with the frequency of electromagnetic radiation can be seen on any sunny day: it is the reason the sky is blue and the setting Sun is red. The higher-frequency blue light from the Sun is scattered much more by the atoms and molecules of the Earth's atmosphere than is the lower-frequency red light. Hence the light of the setting Sun, which passes through a thick layer of atmosphere, has much more red than yellow or blue light, while light scattered from the sky contains much more blue than yellow or red light.

The process of scattering, or reradiating part of the electromagnetic wave by a charge oscillator, is fundamental to understanding the interaction of electromagnetic radiation with solids, liquids, or any matter that contains a very large number of charges and thus an enormous number of charge oscillators. This also explains why a substance that has charge oscillators of certain frequencies absorbs and emits radiation of those frequencies.

Significance of the scattering process

When electromagnetic radiation falls on a large collection of individual small charge oscillators, as in a piece of glass or metal or a brick wall, all of these oscillators perform oscillations in unison, following the beat of the electric wave. As a result, all the oscillators emit secondary radiation in unison (or coherently), and the total secondary radiation coming from the solid consists of the sum of all these secondary coherent electromagnetic waves. This sum total yields radiation that is reflected from the surface of the solid and radiation that goes into the solid at a certain angle with respect to the normal of (*i.e.*, a line perpendicular to) the surface. The latter is the refracted radiation that may be attenuated (absorbed) on its way through the solid.

Superposition and interference. When two electromagnetic waves of the same frequency superpose in space, the resultant electric and magnetic field strength of any point of space and time is the sum of the respective fields of the two waves. When one forms the sum, both the magnitude and the direction of the fields need be considered, which means that they sum like vectors. In the special case when two equally strong waves have their fields in the same direction in space and time (*i.e.*, when they are in phase), the resultant field is twice that of each individual wave. The resultant intensity, being proportional to the square of the field strength, is therefore not two but four times the intensity of each of the two superposing waves.

By contrast, the superposition of a wave that has an electric field in one direction (positive) in space and time with a wave of the same frequency having an electric field in the opposite direction (negative) in space and time leads to cancellation and no resultant wave at all (zero intensity). Two waves of this sort are termed out of phase. The first example, that of in-phase superposition yielding four times the individual intensity, constitutes what is called constructive interference. The second example, that of out-of-phase superposition yielding zero intensity, is destructive interference. Since the resultant field at any point and time is the sum of all individual fields at that point and time, these arguments are easily extended to any number of superposing waves. One finds constructive, destructive, or partial interference for waves having the same frequency and given phase relationships.

Constructive and destructive interference

Propagation and coherence. Once generated, an electromagnetic wave is self-propagating because a time-varying electric field produces a time-varying magnetic field and vice versa. When an oscillating current in an antenna is switched on for, say, eight minutes, then the beginning of the electromagnetic train reaches the Sun just when the antenna is switched off because it takes a few seconds more than eight minutes for electromagnetic radiation to reach the Sun. This eight-minute wave train, which is as long as the Sun–Earth distance, then continues to travel with the speed of light past the Sun into the space beyond.

Except for radio waves transmitted by antennas that are switched on for many hours, most electromagnetic waves comes in many small pieces. The length and duration of a wave train are called coherence length and coherence time, respectively. Light from the Sun or from a light bulb comes in many tiny bursts lasting about a millionth of a millionth of a second and having a coherence length of about one centimetre. The discrete radiant energy emitted by an atom as it changes its internal energy can have a coherence length several hundred times longer (one to 10 metres) unless the radiating atom is disturbed by a collision.

Coherence length and coherence time

The time and space at which the electric and magnetic fields have a maximum value or are zero between the reversal of their directions are different for different wave trains. It is therefore clear that the phenomenon of interference can arise only from the superposition of part of a wave train with itself. This can be accomplished, for instance, with a half-transparent mirror that reflects half the intensity and transmits the other half of each of the billion billion wave trains of a given light source, say, a yellow sodium discharge lamp. One can allow one of these half beams to travel in direction A and the other in direction B, as shown in Figure 4. By reflecting each half beam back, one can then superpose the two half beams and observe the resultant total. If one half beam has to travel a path ½ wavelength or ³⁄₂ or ⁵⁄₂ wavelength longer than the other, then the superposition yields no light at all because the electric and magnetic fields of every half wave train in the two half beams point in opposite directions and their sum is therefore zero. The important point is that cancellation occurs between each half wave train and its mate. This is an example of destructive interference. By adjusting the path lengths A and B such that they are equal or differ by λ, 2λ, $3\lambda \ldots$, the electric and magnetic fields of each half wave train and its mate add when they are superposed. This is constructive interference, and, as a result, one sees strong light.

The interferometer discussed above and represented in Figure 4 was designed by the American physicist Albert A. Michelson in 1880 (while he was studying with Hermann von Helmholtz in Berlin) for the purpose of measuring the effect on the speed of light of the motion of the ether through which light was believed to travel (see below *The electromagnetic wave and field concept*).

Speed of electromagnetic radiation and the Doppler effect. Electromagnetic radiation, or in modern terminology the photons $h\nu$, always travel in free space with the universal speed c—*i.e.*, the speed of light. This is actually a very puzzling situation which was first experimentally verified by Michelson and Edward Williams Morley, another American scientist, in 1887 and which is the basic

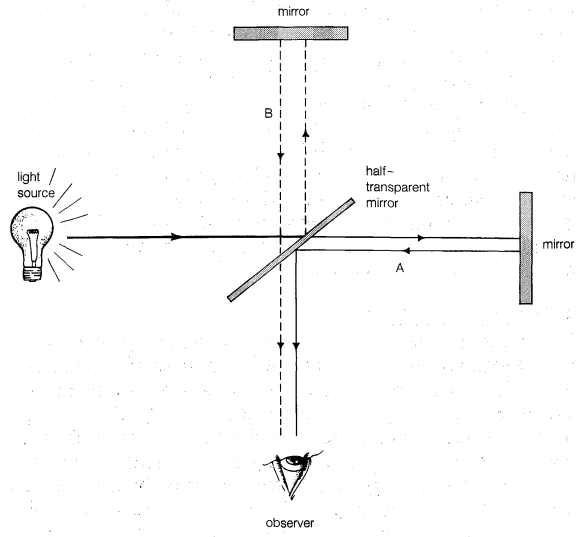


Figure 4: Michelson interferometer.

axiom of Albert Einstein's theory of relativity. Although there is no doubt that it is true, the situation is puzzling because it is so different from the behaviour of normal particles; that is to say, for little or not so little pieces of matter. When one chases behind a normal particle (*e.g.*, an airplane) or moves in the opposite direction toward it, one certainly will measure very different speeds of the airplane relative to oneself. One would detect a very small relative speed in the first case and a very large one in the second. Moreover, a bullet shot forward from the airplane and another toward the back would appear to be moving with different speeds relative to oneself. This would not at all be the case when one measures the speed of electromagnetic radiation: irrespective of one's motion or that of the source of the electromagnetic radiation, any measurement by a moving observer will result in the universal speed of light. This must be accepted as a fact of nature.

What happens to pitch or frequency when the source is moving toward the observer or away from him? It has been established from sound waves that the frequency is higher when a sound source is moving toward the observer and lower when it is moving away from him. This is the Doppler effect, named after the Austrian physicist Christian Doppler, who first described the phenomenon in 1842. Doppler predicted that the effect also occurs with electromagnetic radiation and suggested that it be used for measuring the relative speeds of stars. This means that a characteristic blue light emitted, for example, by an excited helium atom as it changes from a higher to a lower internal energy state would no longer appear blue when one looks at this light coming from helium atoms that move very rapidly away from the Earth with, say, a galaxy. When the speed of such a galaxy away from the Earth is large, the light may appear yellow; if the speed is still larger, it may appear red or even infrared. This is actually what happens, and the speed of galaxies as well as of stars relative to the Earth is measured from the Doppler shift of characteristic atomic radiation energies $h\nu$.

COSMIC BACKGROUND ELECTROMAGNETIC RADIATION

As one measures the relative speeds of galaxies using the Doppler shift of characteristic radiation emissions, one finds that all galaxies are moving away from one another. Those that are moving the fastest are systems that are the farthest away (Hubble's law). The speeds and distances give the appearance of an explosion. Extrapolating backward in time, one obtains an estimate as to when this explosion, dubbed the big bang, might have occurred. This time is calculated to be somewhere between 15 and 20 billion years ago, which is considered to be the age of the universe. From this early stage onward, the universe expanded and cooled. The American scientists Robert W. Wilson and Arno Penzias determined in 1965 that the whole universe can be conceived of as an expanding blackbody filled with electromagnetic radiation which now

The universe as an expanding blackbody

corresponds to a temperature of 2.74 K, only a few degrees above absolute zero. Because of this low temperature, most of the radiation energy is in the microwave region of the electromagnetic spectrum. The intensity of this radiation corresponds, on average, to about 400 photons in every cubic centimetre of the universe. It has been estimated that there are about one billion times more photons in the universe than electrons, nuclei, and all other things taken together. The presence of this microwave cosmic background radiation supports the predictions of big-bang cosmology. (For more specific information on these and related matters, see COSMOS, THE.)

EFFECT OF GRAVITATION

The energy of the quanta of electromagnetic radiation is subject to gravitational forces just like a mass of magnitude $m = h\nu/c^2$. This is so because the relationship of energy E and mass m is $E = mc^2$. As a consequence, light traveling toward the Earth gains energy and its frequency is shifted toward the blue (shorter wavelengths), whereas light traveling "up" loses energy and its frequency is shifted toward the red (longer wavelengths). These shifts are very small but have been detected by the American physicists Robert V. Pound and Glen A. Rebka.

The effect of gravitation on light increases with the strength of the gravitational attraction. Thus, a light beam from a distant star does not travel along a straight line when passing a star like the Sun but is deflected toward it. This deflection can be strong around very heavy cosmic objects, which then distort the light path acting as a gravitational lens.

Under extreme conditions the gravitational force of a cosmic object can be so strong that no electromagnetic radiation can escape the gravitational pull. Such an object, called a black hole, is therefore not visible and its presence can only be detected by its gravitational effect on other visible objects in its vicinity. (For additional information, see COSMOS, THE; PHYSICAL SCIENCES, THE: *Astronomy.*)

THE GREENHOUSE EFFECT OF THE ATMOSPHERE

The temperature of the terrestrial surface environment is controlled not only by the Sun's electromagnetic radiation but also in a sensitive way by the Earth's atmosphere. As noted earlier, each substance absorbs and emits electromagnetic radiation of some energies $h\nu$ and does not do so in other ranges of energy. These regions of transparency and opaqueness are governed by the particular distribution of internal energies of the substance.

The Earth's atmosphere acts much like the glass panes of a greenhouse: it allows sunlight, particularly its visible range, to reach and warm the Earth, but it largely inhibits the infrared radiation emitted by the heated terrestrial surface from escaping into space. Figure 5 shows the absorption of the Earth's atmosphere for various frequencies and wavelengths of electromagnetic radiation. Since the atmosphere becomes thinner and thinner with increasing altitude above the Earth, there is less atmospheric absorption in the higher regions of the atmosphere. At an altitude of 100 kilometres, the fraction of atmosphere is one 10-millionth of that on the ground. Figure 5 shows the altitude at which the intensity of electromagnetic radiation of certain frequencies coming from space is attenuated to one-half of its original value. There are regions of strong absorption and "windows" of transmission. Below 10 million hertz (10^7 Hz), the absorption is caused by the ionosphere, a layer in which atoms and molecules in the atmosphere are ionized by the Sun's ultraviolet radiation. In the infrared region, the absorption is caused by molecular vibrations and rotations. In the ultraviolet and X-ray regions, the absorption is due to electronic excitations in atoms and molecules. The window of transmission for visible light can be seen near the centre of the diagram.

Infrared-absorbing gases in the atmosphere

Without water vapour and carbon dioxide (CO_2), which are, together with certain industrial pollutants, the main infrared-absorbing species in the atmosphere, the Earth would experience the extreme temperature variations between night and day that occur on the Moon. The Earth would then be a frozen planet, like Mars, with an average temperature of 200 K, and not be able to support life.

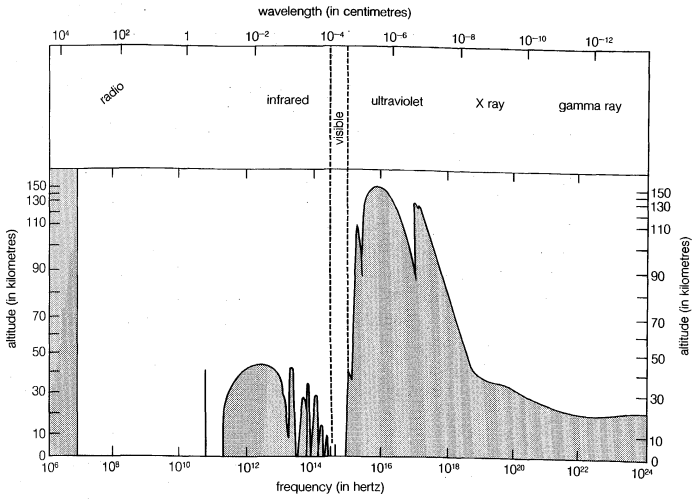


Figure 5: *Atmospheric absorption of electromagnetic radiation.* The intensity of electromagnetic radiation of certain frequencies coming from space is attenuated by the Earth's atmosphere to half of its original value at the altitudes shown here.

Adapted from "The New Astrophysics" by M. Longair in P. Davies (ed.), *The New Physics*, © Cambridge University Press, 1989; after R. Giacconi, H. Gurksy, and L.P. van Speybroeck in "Observational Techniques in X-Ray Astronomy," reproduced, with permission, from the *Annual Review of Astronomy and Astrophysics*, vol. 6, © 1968 by Annual Reviews Inc.

Scientists believe that the Earth's temperature and climate in general will be affected as the composition of the atmosphere is altered by an increased release and accumulation of carbon dioxide and other gaseous pollutants (for a detailed discussion, see CLIMATE AND WEATHER: *Climate and life: Impact of human activities on climate;* HYDROSPHERE, THE: *Impact of human activities on the hydrosphere: Buildup of greenhouse gases*).

Forms of electromagnetic radiation

Electromagnetic radiation appears in a wide variety of forms and manifestations. Yet, these diverse phenomena are understood to comprise a single aspect of nature, following simple physical principles. Common to all forms is the fact that electromagnetic radiation interacts with and is generated by electric charges. The apparent differences in the phenomena arise from the question in which environment and under what circumstances can charges respond on the time scale of the frequency ν of the radiation.

At smaller frequencies ν (smaller than 10^{12} hertz), electric charges typically are the freely moving electrons in the metal components of antennas or the free electrons and ions in space that give rise to phenomena related to radio waves, radar waves, and microwaves. At higher frequencies (10^{12} to 5×10^{14} hertz), in the infrared region of the spectrum, the moving charges are primarily associated with the rotations and vibrations of molecules and the motions of atoms bonded together in materials. Electromagnetic radiation in the visible range to X rays have frequencies that correspond to charges within atoms, whereas gamma rays are associated with frequencies of charges within atomic nuclei. The characteristics of electromagnetic radiation occurring in the different regions of the spectrum are described in this section.

RADIO WAVES

Radio waves are used for wireless transmission of sound messages, or information, for communication, as well as for maritime and aircraft navigation. The information is imposed on the electromagnetic carrier wave as amplitude modulation (AM) or as frequency modulation (FM) or in digital form (pulse modulation). Transmission therefore involves not a single-frequency electromagnetic wave but rather a frequency band whose width is proportional to the information density. The width is about 10,000 Hz for telephone, 20,000 Hz for high-fidelity sound, and five megahertz (MHz = one million hertz) for high-definition television. This width and the decrease in efficiency of generating electromagnetic waves with decreasing fre-

quency sets a lower frequency limit for radio waves near 10,000 Hz.

Because electromagnetic radiation travels in free space in straight lines, scientists questioned the efforts of the Italian physicist and inventor Guglielmo Marconi to develop long-range radio. The curvature of the Earth limits the line-of-sight distance from the top of a 100-metre (330-foot) tower to about 30 kilometres (19 miles). Marconi's unexpected success in transmitting messages over more than 2,000 kilometres led to the discovery of the Kennelly–Heaviside layer, more commonly known as the ionosphere. This region is an approximately 300-kilometre-thick layer starting about 100 kilometres above the Earth's surface in which the atmosphere is partially ionized by ultraviolet light from the Sun, giving rise to enough electrons and ions to affect radio waves. Because of the Sun's involvement, the height, width, and degree of ionization of the stratified ionosphere vary from day to night and from summer to winter.

Radio waves transmitted by antennas in certain directions are bent or even reflected back to Earth by the ionosphere, as illustrated in Figure 6. They may bounce off the Earth and be reflected by the ionosphere repeatedly, making radio transmission around the globe possible. Long-distance communication is further facilitated by the so-called ground wave. This form of electromagnetic wave closely follows the surface of the Earth, particularly over water, as a result of the wave's interaction with the terrestrial surface. The range of the ground wave (up to 1,600 kilometres) and the bending and reflection of the sky wave by the ionosphere depend on the frequency of the waves. Under normal ionospheric conditions 40 MHz is the highest-frequency radio wave that can be reflected from the ionosphere. The strong absorption of the ionosphere below 10 MHz is shown in Figure 5. In order to accommodate the large band width of transmitted signals, television frequencies are necessarily higher than 40 MHz. Television transmitters must therefore be placed on high towers or on hilltops.

Ground waves

As a radio wave travels from the transmitting to the receiving antenna, it may be disturbed by reflections from buildings and other large obstacles. Disturbances arise when several such reflected parts of the wave reach the receiving antenna and interfere with the reception of the wave. Radio waves can penetrate nonconducting materials such as wood, bricks, and concrete fairly well.

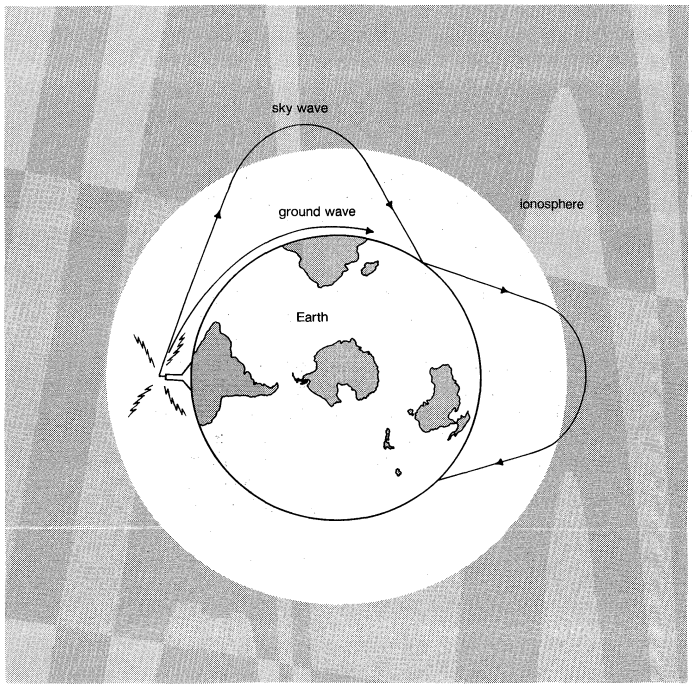


Figure 6: Radio-wave transmission reaching beyond line of sight by means of the sky wave reflected by the ionosphere and by means of the ground wave (see text).

They cannot pass through electrical conductors such as water or metals. Above $\nu = 40$ MHz, radio waves from deep space can penetrate the Earth's atmosphere. This makes radio astronomy observations with ground-based telescopes possible.

Whenever transmission of electromagnetic energy from one location to another is required with minimal energy loss and disturbance, the waves are confined to a limited region by means of wires, coaxial cables, and, in the microwave region, waveguides. Unguided or wireless transmission is naturally preferred when the locations of receivers are unspecified or too numerous, as in the case of radio and television communications. Cable television, as the name implies, is an exception. In this case electromagnetic radiation is transmitted by a coaxial cable system to users either from a community antenna or directly from broadcasting stations. The shielding of this guided transmission from disturbances provides high-quality signals.

Transmission via wires, coaxial cables, and waveguides

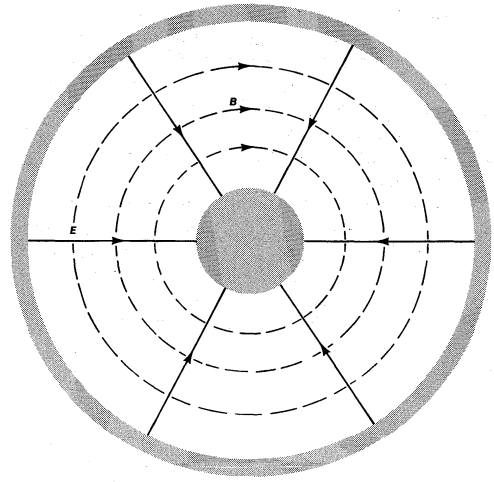


Figure 7: *Cross section of a coaxial cable carrying high-frequency current.*
Electric field lines ***E*** (solid) and magnetic field lines ***B*** (dashed) are mutually perpendicular and perpendicular to the electromagnetic wave propagation, which is toward the viewer.

Figure 7 shows the electric field ***E*** (solid lines) and the magnetic field ***B*** (dashed lines) of an electromagnetic wave guided by a coaxial cable. There is a potential difference between the inner and outer conductors and so electric field lines ***E*** extend from one conductor to the other, represented here in cross section. The conductors carry opposite currents that produce the magnetic field lines ***B***. The electric and magnetic fields are perpendicular to each other and perpendicular to the direction of propagation, as is characteristic of the electromagnetic waves illustrated in Figure 2. At any cross section viewed, the directions of the ***E*** and ***B*** field lines change to their opposite with the frequency ν of the radiation. This direction reversal of the fields does not change the direction of propagation along the conductors. The speed of propagation is again the universal speed of light if the region between the conductors consists of air or free space.

A combination of radio waves and strong magnetic fields is used by magnetic resonance imaging (MRI) to produce diagnostic pictures of parts of the human body and brain without apparent harmful effects. This imaging technique has thus found increasingly wider application in medicine (see also RADIATION: *Imaging techniques*).

Extremely low-frequency (ELF) waves are of interest for communications systems for submarines. The relatively weak absorption by seawater of electromagnetic radiation at low frequencies and the existence of prominent resonances of the natural cavity formed by the Earth and the ionosphere make the range between 5 and 100 Hz attractive for this application.

There is evidence that ELF waves and the oscillating magnetic fields that occur near electric power transmission lines or electric heating blankets have adverse effects on human health and the electrochemical balance of the

brain. Prolonged exposure to low-level and low-frequency magnetic fields have been reported to increase the risk of developing leukemia, lymphoma, and brain cancer in children.

MICROWAVES

The microwave region extends from 1,000 to 300,000 MHz (or 30-centimetre to one-millimetre wavelengths). Although microwaves were first produced and studied in 1886 by Hertz, their practical application had to await the invention of suitable generators, such as the klystron and magnetron.

Use in high-speed telegraphic data transmissions

Microwaves are the principal carriers of high-speed telegraphic data transmissions between stations on the Earth and also between ground-based stations and satellites and space probes. A system of synchronous satellites about 36,000 kilometres above the Earth is used for international broadband telegraphy of all kinds of communications—*e.g.*, television, telephone, and telefacsimile (FAX).

Microwave transmitters and receivers are parabolic dish antennas. They produce microwave beams whose spreading angle is proportional to the ratio of the wavelength of the constituent waves to the diameter of the dish. The beams can thus be directed like a searchlight. Radar beams consist of short pulses of microwaves. One can determine the distance of an airplane or ship by measuring the time it takes such a pulse to travel to the object and, after reflection, back to the radar dish antenna. Moreover, by making use of the change in frequency of the reflected wave pulse caused by the Doppler effect (see above), one can measure the speed of objects. Microwave radar is therefore widely used for guiding airplanes and vessels and for detecting speeding motorists. Microwaves can penetrate clouds of smoke, but are scattered by water droplets, and so are used for mapping meteorologic disturbances and in weather forecasting (see CLIMATE AND WEATHER: *Meteorological measurement and weather forecasting*).

Microwaves play an increasingly wide role in heating and cooking food. They are absorbed by water and fat in foodstuffs (*e.g.*, in the tissue of meats) and produce heat from the inside. In most cases, this reduces the cooking time a hundredfold. Such dry objects as glass and ceramics, on the other hand, are not heated in the process, and metal foils are not penetrated at all.

The heating effect of microwaves destroys living tissue when the temperature of the tissue exceeds 43° C (109° F). Accordingly, exposure to intense microwaves in excess of 20 milliwatts of power per square centimetre of body surface is harmful. The lens of the human eye is particularly affected by waves with a frequency of 3,000 MHz, and repeated and extended exposure can result in cataracts. Radio waves and microwaves of far less power (microwatts per square centimetre) than the 10–20 milliwatts per square centimetre needed to produce heating in living tissue can have adverse effects on the electrochemical balance of the brain and the development of a fetus if these waves are modulated or pulsed at low frequencies between 5 and 100 hertz, which are of the same magnitude as brain wave frequencies.

Microwave sources

Various types of microwave generators and amplifiers have been developed. Vacuum-tube devices, the klystron and the magnetron, continue to be used on a wide scale, especially for higher-power applications. Klystrons are primarily employed as amplifiers in radio relay systems and for dielectric heating, while magnetrons have been adopted for radar systems and microwave ovens. (For a detailed discussion of these devices, see ELECTRONICS: *Principal devices and components: Electron tubes.*) Solid-state technology has yielded several devices capable of producing, amplifying, detecting, and controlling microwaves. Notable among these are the Gunn diode and the tunnel (or Esaki) diode. Another type of device, the maser (acronym for "microwave amplification by stimulated emission of radiation") has proved useful in such areas as radio astronomy, microwave radiometry, and long-distance communications.

Astronomers have discovered what appears to be natural masers in some interstellar clouds. Observations of radio radiation from interstellar hydrogen (H_2) and certain other molecules indicate amplification by the maser process. Also, as was mentioned above, microwave cosmic background radiation has been detected and is considered by many to be the remnant of the primeval fireball postulated by the big-bang cosmological model.

INFRARED RADIATION

Beyond the red end of the visible range but at frequencies higher than those of radar waves and microwaves is the infrared region of the electromagnetic spectrum, between frequencies of 10^{12} and 5×10^{14} Hz (or wavelengths from 0.1 to 7.5×10^{-5} centimetre). William Herschel, a German-born British musician and self-taught astronomer, discovered this form of radiation in 1800 by exploring, with the aid of a thermometer, sunlight dispersed into its colours by a glass prism. Infrared radiation is absorbed and emitted by the rotations and vibrations of chemically bonded atoms or groups of atoms and thus by many kinds of materials. For instance, window glass that is transparent to visible light absorbs infrared radiation by the vibration of its constituent atoms. Infrared radiation is strongly absorbed by water and by the atmosphere, as shown in Figures 3 and 5, respectively. Although invisible to the eye, infrared radiation can be detected as warmth by the skin. Nearly 50 percent of the Sun's radiant energy is emitted in the infrared region of the electromagnetic spectrum, with the rest primarily in the visible region.

Atmospheric haze and certain pollutants that scatter visible light are nearly transparent to parts of the infrared spectrum because the scattering efficiency increases with the fourth power of the frequency. Infrared photography of distant objects from the air takes advantage of this phenomenon. For the same reason, infrared astronomy enables researchers to observe cosmic objects through large clouds of interstellar dust that scatter infrared radiation substantially less than visible light. However, since water vapour, ozone, and carbon dioxide in the atmosphere absorb large parts of the infrared spectrum most infrared astronomical observations are carried out at high altitude by balloons, rockets, or spacecraft.

An infrared photograph of a landscape enhances objects according to their heat emission: blue sky and water appear nearly black, whereas green foliage and unexposed skin show up brightly. Infrared photography can reveal pathological tissue growths (thermography) and defects in electronic systems and circuits due to their increased emission of heat.

The infrared absorption and emission characteristics of molecules and materials yield important information about the size, shape, and chemical bonding of molecules and of atoms and ions in solids. The energies of rotation and vibration are quantized in all systems. The infrared radiation energy $h\nu$ emitted or absorbed by a given molecule or substance is therefore a measure of the difference of some of the internal energy states. These in turn are determined by the atomic weight and molecular bonding forces. For this reason, infrared spectroscopy is a powerful tool for determining the internal structure of molecules and substances or, when such information is already known and tabulated, for identifying the amounts of those species in a given sample. Infrared spectroscopic techniques are often used to determine the composition and hence the origin and age of archaeological specimens and for detecting forgeries of art and other objects, which, when inspected under visible light, resemble the originals.

Infrared radiation plays an important role in heat transfer and is integral to the so-called greenhouse effect (see above), influencing the thermal radiation budget of the Earth on a global scale and affecting nearly all biospheric activity. Virtually every object at the Earth's surface emits electromagnetic radiation primarily in the infrared region of the spectrum.

Infrared LEDs and lasers

Man-made sources of infrared radiation include, besides hot objects, infrared light-emitting diodes (LEDs) and lasers. LEDs are small, inexpensive optoelectronic devices made of such semiconducting materials as gallium arsenide. Infrared LEDs are employed as optoisolators and as light sources in some fibre-optics-based communications systems (see ELECTRONICS: *Principal devices and compo-*

nents: Optoelectronic devices). Powerful optically pumped infrared lasers have been developed using carbon dioxide and carbon monoxide. Carbon dioxide infrared lasers are used to induce and alter chemical reactions and in isotope separation. They also are employed in LIDAR (light radar) systems. Other applications of infrared light include its use in the rangefinders of automatic self-focusing cameras, security alarm systems, and night-vision optical instruments.

Instruments for detecting infrared radiation include heat-sensitive devices such as thermocouple detectors, bolometers (some of these are cooled to temperatures close to absolute zero so that the thermal radiation of the detector system itself is greatly reduced), photovoltaic cells, and photoconductors. The latter are made of semiconductor materials (*e.g.,* silicon and lead sulfide) whose electrical conductance increases when exposed to infrared radiation.

VISIBLE RADIATION

Visible light is the most familiar form of electromagnetic radiation and makes up that portion of the spectrum to which the eye is sensitive. This span is very narrow; the frequencies of violet light are only about twice those of red. The corresponding wavelengths extend from 7×10^{-5} centimetre (red) to 4×10^{-5} centimetre (violet). The energy of a photon from the centre of the visible spectrum (yellow) is $h\nu = 2.2$ eV. This is one million times larger than the energy of a photon of a television wave and one billion times larger than that of radio waves in general (see Figure 1).

Life on Earth could not exist without visible light, which represents the peak of the Sun's spectrum and close to one-half of all of its radiant energy. Visible light is essential for photosynthesis, which enables plants to produce the carbohydrates and proteins that are the food sources for animals. Coal and oil are sources of energy accumulated from sunlight in plants and microorganisms millions of years ago, and hydroelectric power is extracted from one step of the hydrologic cycle kept in motion by sunlight at the present time.

Considering the importance of visible sunlight for all aspects of terrestrial life, one cannot help being awed by the dramatically narrow window in the atmospheric absorption shown in Figure 5 and in the absorption spectrum of water in Figure 3. The remarkable transparency of water centred in the narrow regime of visible light, indicated by vertical dashed lines in Figure 3, is the result of the characteristic distribution of internal energy states of water. Absorption is strong toward the infrared on account of molecular vibrations and intermolecular oscillations. In the ultraviolet region, absorption of radiation is caused by electronic excitations. Light of frequencies having absorption coefficients larger than $\alpha = 10$ cm^{-1} cannot even reach the retina of the human eye because its constituent liquid consists mainly of water that absorbs such frequencies of light.

Conversion of solar energy into electricity

Since the 1970s an increasing number of devices have been developed for converting sunlight into electricity. Unlike various conventional energy sources, solar energy does not become depleted by use and does not pollute the environment. Two branches of development may be noted—namely, photothermal and photovoltaic technologies. In photothermal devices, sunlight is used to heat a substance, as, for example, water, to produce steam with which to drive a generator. Photovoltaic devices, on the other hand, convert the energy in sunlight directly to electricity by use of the photovoltaic effect in a semiconductor junction. Solar panels consisting of photovoltaic devices made of gallium arsenide have conversion efficiencies of more than 20 percent and are used to provide electric power in many satellites and space probes. Large-area solar panels can be made with amorphous semiconductors that have conversion efficiencies of about 10 percent. Solar cells have replaced dry cell batteries in some portable electronic instruments, and solar energy power stations of one- to six-megawatts capacity have been built.

The intensity and spectral composition of visible light can be measured and recorded by essentially any process or property that is affected by light. Detectors make use of a photographic process based on silver halide, the photoemission of electrons from metal surfaces, the generation of electric current in a photovoltaic cell, and the increase in electrical conduction in semiconductors.

Glass fibres constitute an effective means of guiding and transmitting light. A beam of light is confined by total internal reflection to travel inside such an optical fibre, whose thickness may be anywhere between one hundredth of a millimetre and a few millimetres. Many thin optical fibres can be combined into bundles to achieve image reproduction. The flexibility of these fibres or fibre bundles permits their use in medicine for optical exploration of internal organs. Optical fibres connecting the continents provide the capability to transmit substantially larger amounts of information than other systems of international telecommunications. Another advantage of optical fibre communication systems is that transmissions cannot easily be intercepted and are not disturbed by lower atmospheric and stratospheric disturbances.

Optical fibres integrated with miniature semiconductor lasers and light-emitting diodes, as well as with light detector arrays and photoelectronic imaging and recording materials, form the building blocks of a new optoelectronics industry. Some familiar commercial products are optoelectronic copying machines, laser printers, compact disc players, FAX machines, optical recording media, and optical disc mass-storage systems of exceedingly high bit density.

ULTRAVIOLET RADIATION

The German physicist Johann Wilhelm Ritter, having learned of Herschel's discovery of infrared waves, looked beyond the violet end of the visible spectrum of the Sun and found (in 1801) that there exist invisible rays that darken silver chloride even more efficiently than visible light. This spectral region extending between visible light and X rays is designated ultraviolet. Sources of this form of electromagnetic radiation are hot objects like the Sun, synchrotron radiation sources, mercury or xenon arc lamps, and gaseous discharge tubes filled with gas atoms (*e.g.,* mercury, deuterium, or hydrogen) that have internal electron energy levels which correspond to the photons of ultraviolet light.

Fluorescence

When ultraviolet light strikes certain materials, it causes them to fluoresce—*i.e.,* they emit electromagnetic radiation of lower energy, such as visible light. The spectrum of fluorescent light is characteristic of a material's composition and thus can be used for screening minerals, detecting bacteria in spoiled food, identifying pigments, or detecting forgeries of artworks and other objects (the aged surfaces of ancient marble sculptures, for instance, fluoresce yellow-green, whereas a freshly cut marble surface fluoresces bright violet).

Optical instruments for the ultraviolet region are made of special materials, such as quartz, certain silicates, and metal fluorides, which are transparent at least in the near ultraviolet. Far-ultraviolet radiation is absorbed by nearly all gases and materials and thus requires reflection optics in vacuum chambers.

Ultraviolet radiation is detected by photographic plates and by means of the photoelectric effect in photomultiplier tubes. Also, ultraviolet radiation can be converted to visible light by fluorescence before detection.

The relatively high energy of ultraviolet light gives rise to certain photochemical reactions. This characteristic is exploited to produce cyanotype impressions on fabrics and for blueprinting design drawings. Here, the fabric or paper is treated with a mixture of chemicals that react upon exposure to ultraviolet light to form an insoluble blue compound. Electronic excitations caused by ultraviolet radiation also produce changes in the colour and transparency of photosensitive and photochromic glasses. Photochemical and photostructural changes in certain polymers constitute the basis for photolithography and the processing of the microelectronic circuits.

Although invisible to the eyes of humans and most vertebrates, near-ultraviolet light can be seen by many insects. Butterflies and many flowers that appear to have identical colour patterns under visible light are distinctly

different when viewed under the ultraviolet rays perceptible to insects.

An important difference between ultraviolet light and electromagnetic radiation of lower frequencies is the ability of the former to ionize, meaning that it can knock an electron out from atoms and molecules. All high-frequency electromagnetic radiation beyond the visible—*i.e.*, ultraviolet light, X rays, and gamma rays—is ionizing and therefore harmful to body tissues, living cells, and DNA (deoxyribonucleic acid). The harmful effects of ultraviolet light to humans and larger animals are mitigated by the fact that this form of radiation does not penetrate much further than the skin.

The body of a sunbather is struck by 10^{21} photons every second, and 1 percent of these, or more than a billion billion per second, are photons of ultraviolet radiation. Tanning and natural body pigments help to protect the skin to some degree, preventing the destruction of skin cells by ultraviolet light. Nevertheless, overexposure to the ultraviolet component of sunlight can cause skin cancer, cataracts of the eyes, and damage to the body's immune system. Fortunately a layer of ozone (O_3) in the stratosphere absorbs the most damaging ultraviolet rays, which have wavelengths of 2000 and 2900 angstroms (one angstrom [Å] = 10^{-10} metre), and attenuates those with wavelengths between 2900 and 3150 Å, as shown in Figure 5. Without this protective layer of ozone, life on Earth would not be possible. The ozone layer is produced at an altitude of about 10 to 50 kilometres above the Earth's surface by a reaction between upward-diffusing molecular oxygen (O_2) and downward-diffusing ionized atomic oxygen (O^+). Many scientists believe that this life-protecting stratospheric ozone layer is being reduced by chlorine atoms in chlorofluorocarbon (or Freon) gases released into the atmosphere by aerosol propellants, air-conditioner coolants, solvents used in the manufacture of electronic components, and other sources. (For more specific information, see ATMOSPHERE: *Composition of the present atmosphere: Effects of human activity on atmospheric composition and their ramifications: Depletion of stratospheric ozone.*)

Protective ozone layer

Ionized atomic oxygen, nitrogen, and nitric oxide are produced in the upper atmosphere by absorption of solar ultraviolet radiation. This ionized region is the ionosphere, which affects radio communications and reflects and absorbs radio waves of frequencies below 40 MHz (see Figure 5).

X RAYS

The German physicist Wilhelm Conrad Röntgen discovered X rays in 1895 by accident while studying cathode rays in a low-pressure gas discharge tube. (A few years later J.J. Thomson of England showed that cathode rays were electrons emitted from the negative electrode [cathode] of the discharge tube.) Röntgen noticed the fluorescence of a barium platinocyanide screen that happened to lie near the discharge tube. He traced the source of the hitherto undetected form of radiation to the point where the cathode rays hit the wall of the discharge tube, and mistakenly concluded from his inability to observe reflection or refraction that his new rays were unrelated to light. Because of his uncertainty about their nature, he called them X-radiation. This early failure can be attributed to the very short wavelengths of X rays (10^{-8} to 10^{-11} centimetre), which correspond to photon energies from 200 to 100,000 eV. In 1912 another German physicist, Max von Laue, realized that the regular arrangement of atoms in crystals should provide a natural grating of the right spacing (about 10^{-8} centimetre) to produce an interference pattern on a photographic plate when X rays pass through such a crystal. The success of this experiment, carried out by Walter Friedrich and Paul Knipping, not only identified X rays with electromagnetic radiation but also initiated the use of X rays for studying the detailed atomic structure of crystals. The interference of X rays diffracted in certain directions from crystals in so-called X-ray diffractometers, in turn, permits the dissection of X-radiation into its different frequencies, just as a prism diffracts and spreads the various colours of light. The spectral composition and characteristic frequencies of X rays emitted by a given X-ray source can thus be measured. As in optical spectroscopy, the X-ray photons emitted correspond to the differences of the internal electronic energies in atoms and molecules. Because of their much higher energies, however, X-ray photons are associated with the inner-shell electrons close to the atomic nuclei, whereas optical absorption and emission are related to the outermost electrons in atoms or in materials in general. Since the outer electrons are used for chemical bonding while the energies of inner-shell electrons remain essentially unaffected by atomic bonding, the identity and quantity of elements that make up a material are more accurately determined by the emission, absorption, or fluorescence of X rays than of photons of visible or ultraviolet light.

Röntgen's X-radiation

The contrast between body parts in medical X-ray photographs (radiographs) is produced by the different scattering and absorption of X rays by bones and tissues. Within months of Röntgen's discovery of X rays and his first X-ray photograph of his wife's hand, this form of electromagnetic radiation became indispensable in orthopedic and dental medicine. The use of X rays for obtaining images of the body's interior has undergone considerable development over the years and has culminated in the highly sophisticated procedure known as computerized axial tomography (CAT; see RADIATION: *Applications of radiation: Medical applications: Imaging techniques*).

Notwithstanding their usefulness in medical diagnosis, the ability of X rays to ionize atoms and molecules and their penetrating power make them a potential health hazard. Exposure of body cells and tissue to large doses of such ionizing radiation can result in abnormalities in DNA that may lead to cancer and birth defects. (For a detailed treatment of the effects of X rays and other forms of ionizing radiation on human health and the levels of such radiation encountered in daily life, see RADIATION: *Biologic effects of ionizing radiation.*)

X rays are produced in X-ray tubes by the deceleration of energetic electrons (bremsstrahlung) as they hit a metal target or by accelerating electrons moving at relativistic velocities in circular orbits (synchrotron radiation; see above). They are detected by their photochemical action in photographic emulsions or by their ability to ionize gas atoms: every X-ray photon produces a burst of electrons and ions, resulting in a current pulse. By counting the rate of such current pulses per second, the intensity of a flux of X rays can be measured. Instruments used for this purpose are called Geiger counters.

Cosmic X-ray sources

X-ray astronomy has revealed very strong sources of X rays in deep space. In the Milky Way Galaxy, of which the solar system is a part, the most intense sources are certain double star systems in which one of the two stars is thought to be either a compact neutron star or a black hole. The ionized gas of the circling companion star falls by gravitation into the compact star, generating X rays that may be more than 1,000 times as intense as the total amount of light emitted by the Sun. At the moment of their explosion, supernovae emit a good fraction of their energy in a burst of X rays.

GAMMA RAYS

Six years after the discovery of radioactivity (1896) by Henri Becquerel of France, the New Zealand-born British physicist Ernest Rutherford found that three different kinds of radiation are emitted in the decay of radioactive substances; these he called alpha, beta, and gamma rays in sequence of their ability to penetrate matter. The alpha particles were found to be identical with the nuclei of helium atoms and the beta rays were identified as electrons. In 1912 it was shown that the much more penetrating gamma rays have all the properties of very energetic electromagnetic radiation, or photons. Gamma-ray photons are between 10,000 and 10,000,000 times more energetic than the photons of visible light when they originate from radioactive atomic nuclei. Gamma rays with a million million times higher energy make up a very small part of the cosmic rays that reach the Earth from supernovae or from other galaxies. The origin of the most energetic gamma rays is not yet known.

During radioactive decay, an unstable nucleus usually emits alpha particles, electrons, gamma rays, and neutrinos spontaneously. In nuclear fission, the unstable nucleus breaks into fragments, which are themselves complex nuclei, along with such particles as neutrons and protons. The resultant stable nuclei or nuclear fragments are usually in a highly excited state and then reach their low-energy ground state by emitting one or more gamma rays. Such a decay scheme is shown schematically in Figure 8 for the unstable nucleus sodium-24 (^{24}Na). Much of what is known about the internal structure and energies of nuclei has been obtained from the emission or resonant absorption of gamma rays by nuclei. Absorption of gamma rays by nuclei can cause them to eject neutrons or alpha particles or it can even split a nucleus like a bursting bubble in what is called photodisintegration. A gamma particle hitting a hydrogen nucleus (that is, a proton), for example, produces a positive pi-meson and a neutron or a neutral pi-meson and a proton. Neutral pi-mesons, in turn, have a very brief mean life of 1.8×10^{-16} second and decay into two gamma rays of energy $h\nu \approx 70$ MeV. When an energetic gamma ray $h\nu > 1.02$ MeV passes a nucleus, it may disappear while creating an electron–positron pair. Gamma photons interact with matter by discrete elementary processes that include resonant absorption, photodisintegration, ionization, scattering (Compton scattering), or pair production.

Photodisintegration

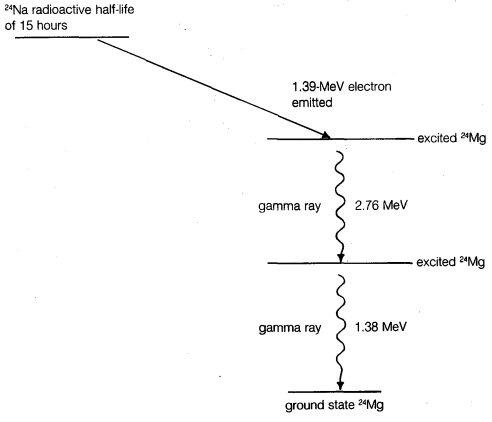


Figure 8: Decay scheme of a radioactive sodium-24 (^{24}Na) nucleus. With a half-life of 15 hours, it decays by beta decay to an excited magnesium-24 (^{24}Mg) nucleus. Two gamma rays are rapidly emitted and the excitation energy is carried off, whereby the stable ground state of magnesium-24 is reached.

Gamma rays are detected by their ability to ionize gas atoms or to create electron–hole pairs in semiconductors or insulators. By counting the rate of charge pulses or voltage pulses or by measuring the scintillation of the light emitted by the subsequently recombining electron–hole pairs, one can determine the number and energy of gamma rays striking an ionization detector or scintillation counter.

Both the specific energy of the gamma-ray photon emitted as well as the half-life of the specific radioactive decay process that yields the photon identify the type of nuclei at hand and their concentrations. By bombarding stable nuclei with neutrons, one can artificially convert more than 70 different stable nuclei into radioactive nuclei and use their characteristic gamma emission for purposes of identification, for impurity analysis of metallurgical specimens (neutron-activation analysis), or as radioactive tracers with which to determine the functions or malfunctions of human organs, to follow the life cycles of organisms, or to determine the effects of chemicals on biological systems and plants.

Penetrating power of gamma rays

The great penetrating power of gamma rays stems from the fact that they have no electric charge and thus do not interact with matter as strongly as do charged particles. Because of their penetrating power gamma rays can be used for radiographing holes and defects in metal castings and other structural parts. At the same time, this property makes gamma rays extremely hazardous. The lethal effect of this form of ionizing radiation makes it useful for sterilizing medical supplies that cannot be sanitized by boiling or for killing organisms that cause food spoilage. More than 50 percent of the ionizing radiation to which humans are exposed comes from natural radon gas, which is an end product of the radioactive decay chain of natural radioactive substances in minerals. Radon escapes from the ground and enters the environment in varying amounts.

Historical survey

DEVELOPMENT OF THE CLASSICAL RADIATION THEORY

The classical electromagnetic radiation theory "remains for all time one of the greatest triumphs of human intellectual endeavor." So said Max Planck in 1931, commemorating the 100th anniversary of the birth of the Scottish physicist James Clerk Maxwell, the prime originator of this theory. The theory was indeed of great significance, for it not only united the phenomena of electricity, magnetism, and light in a unified framework but also was a fundamental revision of the then-accepted Newtonian way of thinking about the forces in the physical universe. The development of the classical radiation theory constituted a conceptual revolution that lasted for nearly half a century. It began with the seminal work of the British physicist and chemist Michael Faraday, who published his article "Thoughts on Ray Vibrations" in *Philosophical Magazine* in May 1846, and came to fruition in 1888 when Hertz succeeded in generating electromagnetic waves at radio and microwave frequencies and measuring their properties.

Wave theory and corpuscular theory. The Newtonian view of the universe may be described as a mechanistic interpretation. All components of the universe, small or large, obey the laws of mechanics, and all phenomena are in the last analysis based on matter in motion. A conceptual difficulty in Newtonian mechanics, however, is the way in which the gravitational force between two massive objects acts over a distance across empty space. Newton did not address this question, but many of his contemporaries hypothesized that the gravitational force was mediated through an invisible and frictionless medium which Aristotle had called the ether (or aether). The problem is that everyday experience of natural phenomena shows mechanical things to be moved by forces which make contact. Any cause and effect without a discernable contact, or "action at a distance," contradicts common sense and has been an unacceptable notion since antiquity. Whenever the nature of the transmission of certain actions and effects over a distance was not yet understood, the ether was resorted to as a conceptual solution of the transmitting medium. By necessity, any description of how the ether functioned remained vague, but its existence was required by common sense and thus not questioned.

Notion of the ether

In Newton's day, light was one phenomenon, besides gravitation, whose effects were apparent at large distances from its source. Newton contributed greatly to the scientific knowledge of light. His experiments revealed that white light is a composite of many colours, which can be dispersed by a prism and reunited to again yield white light. The propagation of light along straight lines convinced him that it consists of tiny particles which emanate at high or infinite speed from the light source. The first observation from which a finite speed of light was deduced was made soon thereafter, in 1676, by the Danish astronomer Ole Rømer (see *Speed of light* below).

Observations of two phenomena strongly suggested that light propagates as waves. One of these involved interference by thin films, which was discovered in England independently by Robert Boyle and Robert Hooke. The other had to do with the diffraction of light in the geometric shadow of an opaque screen. The latter was also discovered by Hooke, who published a wave theory of light in 1665 to explain it.

The Dutch scientist Christiaan Huygens greatly improved the wave theory and explained reflection and refraction in terms of what is now called Huygens' principle. According to this principle (published in 1690), each point on a wave front in the hypothetical ether or in an optical medium is a source of a new spherical light wave and the wave front

Huygens' principle

is the envelope of all the individual wavelets that originate from the old wave front.

In 1669 another Danish scientist, Erasmus Bartholin, discovered the polarization of light by double refraction in Iceland spar (calcite). This finding had a profound effect on the conception of the nature of light. At that time, the only waves known were those of sound, which are longitudinal. It was inconceivable to both Newton and Huygens that light could consist of transverse waves in which vibrations are perpendicular to the direction of propagation. Huygens gave a satisfactory account of double refraction by proposing that the asymmetry of the structure of Iceland spar causes the secondary wavelets to be ellipsoidal instead of spherical in his wave front construction. Since Huygens believed in longitudinal waves, he failed, however, to understand the phenomena associated with polarized light. Newton, on the other hand, used these phenomena as the bases for an additional argument for his corpuscular theory of light. Particles, he argued in 1717, have "sides" and can thus exhibit properties that depend on the directions perpendicular to the direction of motion.

It may be surprising that Huygens did not make use of the phenomenon of interference to support his wave theory; but for him waves were actually pulses instead of periodic waves with a certain wavelength. One should bear in mind that the word wave may have a very different conceptual meaning and convey different images at various times to different people.

It took nearly a century before a new wave theory was formulated by the physicists Thomas Young of England and Augustin-Jean Fresnel of France. Based on his experiments on interference, Young realized for the first time that light is a transverse wave. Fresnel then succeeded in explaining all optical phenomena known at the beginning of the 19th century with a new wave theory. No proponents of the corpuscular light theory remained. Nonetheless, it is always satisfying when a competing theory is discarded on grounds that one of its principal predictions is contradicted by experiment. The corpuscular theory explained the refraction of light passing from a medium of given density to a denser one in terms of the attraction of light particles into the latter. This means the light velocity should be larger in the denser medium. Huygens' construction of wave fronts waving across the boundary between two optical media predicted the opposite—that is to say, a smaller light velocity in the denser medium. The measurement of the light velocity in air and water by Armand-Hippolyte-Louis Fizeau and independently by Jean-Bernard-Léon Foucault during the mid-19th century decided the case in favour of the wave theory (see *Speed of light* below).

The transverse wave nature of light implied that the ether must be a solid elastic medium. The larger velocity of light suggested, moreover, a great elastic stiffness of this medium; yet, it was recognized that all celestial bodies move through the ether without encountering such difficulties as friction. These conceptual problems remained unsolved until the beginning of the 20th century. (Ht.F.)

Relation between electricity and magnetism. As early as 1760 the Swiss-born mathematician Leonhard Euler suggested that the same ether that propagates light is responsible for electrical phenomena. In comparison with both mechanics and optics, however, the science of electricity was slow to develop. Magnetism was the one science that made progress in the Middle Ages, following the introduction from China into the West of the magnetic compass, but electromagnetism played little part in the scientific revolution of the 17th century. It was, however, the only part of physics in which very significant progress was made during the 18th century. By the end of that century the laws of electrostatics—the behaviour of charged particles at rest—were well known, and the stage was set for the development of the elaborate mathematical description first made by the French mathematician Siméon-Denis Poisson. There was no apparent connection of electricity with magnetism, except that magnetic poles, like electric charges, attract and repel with an inverse-square law force.

Following the discoveries in electrochemistry (the chemical effects of electrical current) by the Italian investigators Luigi Galvani, a physiologist, and Alessandro Volta, a physicist, interest turned to current electricity. A search was made by the Danish physicist Hans Christian Ørsted for some connection between electric currents and magnetism, and during the winter of 1819–20 he observed the effect of a current on a magnetic needle. Members of the French Academy learned about Ørsted's discovery in September 1820, and several of them began to investigate it further. Of these, the most thorough in both experiment and theory was the physicist André-Marie Ampère, who may be called the father of electrodynamics. The magnetic effect of a current had been observed earlier (1802) by an Italian jurist, Gian Domenico Romagnosi, but the announcement was published in an obscure newspaper.

The list of four fundamental empirical laws of electricity and magnetism was made complete with the discovery of electromagnetic induction by both Faraday and Joseph Henry in about 1831. In brief, a change in magnetic flux through a conducting circuit produces a current in the circuit. The observation that the induced current is in a direction to oppose the change that produces it, now known as Lenz's law, was formulated by a Russian-born physicist, Heinrich Friedrich Emil Lenz, in 1834. When the laws were put into mathematical form by Maxwell, the law of induction was generalized to include the production of electric force in space, independent of actual conducting circuits, but was otherwise unchanged. On the other hand, Ampère's law describing the magnetic effect of a current required amendment in order to be consistent with the conservation of charge (the total charge must remain constant) in the presence of changing electric fields, and Maxwell introduced the idea of "displacement current" to make the set of equations logically consistent. As a result, he found on combining the equations that he arrived at a wave equation, according to which transverse electric and magnetic disturbances were propagated with a velocity that could be calculated from electrical measurements. These measurements were available to Maxwell, having been made in 1856 by the German physicists Rudolph Hermann Arndt Kohlrausch and Wilhelm Eduard Weber, and his calculation gave him a result that was the same, within the limits of error, as the speed of light in vacuum. It was the coincidence of this value with the velocity of the waves predicted by his theory that convinced Maxwell of the electromagnetic nature of light. (M.Ph./Ht.F.)

Electromagnetic induction

The electromagnetic wave and field concept. Faraday introduced the concept of field and of field lines of force that exist outside material bodies. As he explained it, the region around and outside a magnet or an electric charge contains a field that describes at any location the force experienced by another small magnet or charge placed there. The lines of force around a magnet can be made visible by iron filings sprayed on a paper that is held over the magnet. The concept of field, specifying as it does a certain possible action or force at any location in space, was the key to understanding electromagnetic phenomena. It should be mentioned parenthetically that the field concept also plays (in varied forms) a pivotal role in modern theories of particles and forces.

Besides introducing this important concept of electric and magnetic field lines of force, Faraday had the extraordinary insight that electrical and magnetic actions are not transmitted instantaneously but after a certain lag in time, which increases with distance from the source. Moreover, he realized the connection between magnetism and light after observing that a substance such as glass can rotate the plane of polarization of light in the presence of a magnetic field. This remarkable phenomenon is known as the Faraday effect.

Faraday effect

As noted above, Maxwell formulated a quantitative theory that linked the fundamental phenomena of electricity and magnetism and that predicted electromagnetic waves propagating with a speed, which, as well as one could determine at that time, was identical with the speed of light. He concluded his paper "On the Physical Lines of Force" (1861–62) by saying that electricity may be disseminated through space with properties identical with those of light. In 1864 Maxwell wrote that the numerical factor linking the electrostatic and the magnetic units was very close to

the speed of light and that these results "show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to [his] electromagnetic laws."

What more was needed to convince the scientific community that the mystery of light was solved and the phenomena of electricity and magnetism were unified in a grand theory? Why did it take 25 more years for Maxwell's theory to be accepted? For one, there was little direct proof of the new theory. Furthermore, Maxwell not only had adopted a complicated formalism but also explained its various aspects by unusual mechanical concepts. Even though he stated that all such phrases are to be considered as illustrative and not as explanatory, the French mathematician Henri Poincaré remarked in 1899 that the "complicated structure" which Maxwell attributed to the ether "rendered his system strange and unattractive."

The ideas of Faraday and Maxwell that the field of force has a physical existence in space independent of material media were too new to be accepted without direct proof. On the Continent, particularly in Germany, matters were further complicated by the success of Carl Friedrich Gauss and Wilhelm Eduard Weber in developing a potential field theory for the phenomena of electrostatics and magnetostatics and their continuing effort to extend this formalism to electrodynamics.

Hertz's contributions

It is difficult in hindsight to appreciate the reluctance to accept the Faraday–Maxwell theory. The impasse was finally removed by Hertz's work. In 1884 Hertz derived Maxwell's theory by a new method and put its fundamental equations into their present-day form. In so doing, he clarified the equations, making the symmetry of electric and magnetic fields apparent. The German physicist Arnold Sommerfeld spoke for most of his learned colleagues when, after reading Hertz's paper, he remarked, "the shades fell from my eyes," and admitted that he understood electromagnetic theory for the first time. Four years later, Hertz made a second major contribution: he succeeded in generating electromagnetic radiation of radio and microwave frequencies, measuring their speed by a standing-wave method and proving that these waves have the properties of reflection, diffraction, refraction, and interference common to light. He showed that such electromagnetic waves can be polarized, that the electric and magnetic fields oscillate in directions that are mutually perpendicular and transverse to the direction of motion, and that their velocity is the same as the speed of light, as predicted by Maxwell's theory.

Hertz's ingenious experiments not only settled the theoretical misconceptions in favour of Maxwell's electromagnetic field theory but also opened the way for building transmitters, antennas, coaxial cables, and detectors for radio-frequency electromagnetic radiation. In 1896 Marconi received the first patent for wireless telegraphy, and in 1901 he achieved transatlantic radio communication.

The Faraday–Maxwell–Hertz theory of electromagnetic radiation, which is commonly referred to as Maxwell's theory, makes no reference to a medium in which the electromagnetic waves propagate. A wave of this kind is produced, for example, when a line of charges is moved back and forth along the line. Moving charges represent an electric current. In this back-and-forth motion, the current flows in one direction and then in another. As a consequence of this reversal of current direction, the magnetic field around the current (discovered by Ørsted and Ampère) has to reverse its direction. The time-varying magnetic field produces perpendicular to it a time-varying electric field, as discovered by Faraday (Faraday's law of induction). These time-varying electric and magnetic fields spread out from their source, the oscillating current, at the speed of light in free space. The oscillating current in this discussion is the oscillating current in a transmitting antenna, and the time-varying electric and magnetic fields that are perpendicular to one another propagate at the speed of light and constitute an electromagnetic wave. Its frequency is that of the oscillating charges in the antenna. Once generated, it is self-propagating because a time-varying electric field produces a time-varying magnetic field, and vice versa. Electromagnetic radiation travels through space by itself. The belief in the existence of an ether medium, however, was at the time of Maxwell as strong as at the time of Plato and Aristotle. It was impossible to visualize ether because contradictory properties had to be attributed to it in order to explain the phenomena known at any given time. In his article ETHER in the ninth edition of the *Encyclopædia Britannica,* Maxwell described the vast expanse of the substance, some of it possibly even inside the planets, carried along with them or passing through them as the "water of the sea passes through the meshes of a net when it is towed along by a boat."

If one believes in the ether, it is, of course, of fundamental importance to measure the speed of its motion or the effect of its motion on the speed of light. One does not know the absolute velocity of the ether, but as the Earth moves through its orbit around the Sun there should be a difference in ether velocity along and perpendicular to the Earth's motion equal to its speed. If such is the case, the velocity of light and of any other electromagnetic radiation along and perpendicular to the Earth's motion should, predicted Maxwell, differ by a fraction that is equal to the square of the ratio of the Earth's velocity to that of light. This fraction is one part in 100 million.

Michelson set out to measure this effect and, as noted above, designed for this purpose the interferometer sketched in Figure 4. If it is assumed that the interferometer is turned so that half beam A is oriented parallel to the Earth's motion and half beam B is perpendicular to it, then the idea of using this instrument for measuring the effect of the ether motion is best explained by Michelson's words to his children:

> Two beams of light race against each other, like two swimmers, one struggling upstream and back, while the other, covering the same distance, just crosses the river and returns. The second swimmer will always win, if there is any current in the river.

Michelson–Morley experiment

An improved version of the interferometer, in which each half beam traversed its path eight times before both were reunited for interference, was built in 1887 by Michelson in collaboration with Morley. A heavy sandstone slab holding the interferometer was floated on a pool of mercury to allow rotation without vibration. Michelson and Morley could not detect any difference in the two light velocities parallel and perpendicular to the Earth's motion to an accuracy of one part in four billion. This negative result did not, however, shatter the belief in the existence of an ether because the ether could possibly be dragged along with the Earth and thus be stationary around the Michelson–Morley apparatus. Hertz's formulation of Maxwell's theory made it clear that no medium of any sort was needed for the propagation of electromagnetic radiation. In spite of this, ether-drift experiments continued to be conducted until about the mid-1920s. All such tests confirmed Michelson's negative results, and scientists finally came to accept the idea that no ether medium was needed for electromagnetic radiation.

Speed of light. Much effort has been devoted to measuring the speed of light, beginning with the aforementioned work of Rømer in 1676. Rømer noticed that the orbital period of Jupiter's first moon, Io, is apparently slowed as the Earth and Jupiter move away from each other. The eclipses of Io occur later than expected when Jupiter is at its most remote position. This effect is understandable if light requires a finite time to reach the Earth from Jupiter. From this effect, Rømer calculated the time required for light to travel from the Sun to the Earth as 11 minutes. In 1728 James Bradley, an English astronomer, determined the speed of light from the apparent orbital motion of stars that is produced by the orbital motion of the Earth. He computed the time for light to reach the Earth from the Sun as eight minutes, 12 seconds. The first terrestrial measurements were made in 1849 by Fizeau and a year later by Foucault. Michelson improved on Foucault's method and obtained an accuracy of one part in 100,000.

Any measurement of velocity requires, however, a definition of the measure of length and of time. Current techniques allow a determination of the velocity of electromagnetic radiation to a substantially higher degree of precision than permitted by the unit of length that scien-

tists had applied earlier. In 1983 the value of the speed of light was fixed at exactly 299,792,458 metres per second, and this value was adopted as a new standard. As a consequence, the metre was redefined as the length of the path traveled by light in a vacuum over a time interval of 1/299,792,458 of a second. Furthermore, the second—the international unit of time—has been based on the frequency of electromagnetic radiation emitted by a cesium-133 atom.

DEVELOPMENT OF THE QUANTUM THEORY OF RADIATION

After a long struggle electromagnetic wave theory had triumphed. The Faraday–Maxwell–Hertz theory of electromagnetic radiation seemed to be able to explain all phenomena of light, electricity, and magnetism. The understanding of these phenomena enabled one to produce electromagnetic radiation of many different frequencies which had never been observed before and which opened a world of new opportunities. No one suspected that the conceptional foundations of physics were about to change again.

Radiation laws and Planck's light quanta. The quantum theory of absorption and emission of radiation announced in 1900 by Planck ushered in the era of modern physics. He proposed that all material systems can absorb or give off electromagnetic radiation only in "chunks" of energy, quanta E, and that these are proportional to the frequency of that radiation $E = h\nu$. (The constant of proportionality h is, as noted above, called Planck's constant.)

Planck was led to this radically new insight by trying to explain the puzzling observation of the amount of electromagnetic radiation emitted by a hot body and, in particular, the dependence of the intensity of this incandescent radiation on temperature and on frequency. The quantitative aspects of the incandescent radiation constitute the radiation laws.

The Austrian physicist Josef Stefan found in 1879 that the total radiation energy per unit time emitted by a heated surface per unit area increases as the fourth power of its absolute temperature T (Kelvin scale). This means that the Sun's surface, which is at $T = 6{,}000$ K, radiates per unit area $(6{,}000/300)^4 = 20^4 = 160{,}000$ times more electromagnetic energy than does the same area of the Earth's surface, which is taken to be $T = 300$ K. In 1889 another Austrian physicist, Ludwig Boltzmann, used the second law of thermodynamics to derive this temperature dependence for an ideal substance that emits and absorbs all frequencies. Such an object that absorbs light of all colours looks black, and so was called a blackbody. The Stefan–Boltzmann law is written in quantitative form $W = \sigma T^4$, where W is the radiant energy emitted per second and per unit area and the constant of proportionality is $\sigma = 0.136$ calories per metre^2-second-K^4.

Stefan–Boltzmann law

The wavelength or frequency distribution of blackbody radiation was studied in the 1890s by Wilhelm Wien of Germany. It was his idea to use as a good approximation for the ideal blackbody an oven with a small hole. Any radiation that enters the small hole is scattered and reflected from the inner walls of the oven so often that nearly all incoming radiation is absorbed and the chance of some of it finding its way out of the hole again can be made exceedingly small. The radiation coming out of this hole is then very close to the equilibrium blackbody electromagnetic radiation corresponding to the oven temperature. Wien found that the radiative energy dW per wavelength interval $d\lambda$ has a maximum at a certain wavelength λ_m and that the maximum shifts to shorter wavelengths as the temperature T is increased, as illustrated in Figure 9. He found that the product $\lambda_m T$ is an absolute constant: $\lambda_m T = 0.2898$ centimetre-degree Kelvin.

Wien's law of the shift of the radiative power maximum to higher frequencies as the temperature is raised expresses in a quantitative form commonplace observations. Warm objects emit infrared radiation, which is felt by the skin; near $T = 950$ K a dull red glow can be observed; and the colour brightens to orange and yellow as the temperature is raised. The tungsten filament of a light bulb is $T = 2{,}500$ K hot and emits bright light, yet the peak of its spectrum is still in the infrared according to Wien's law.

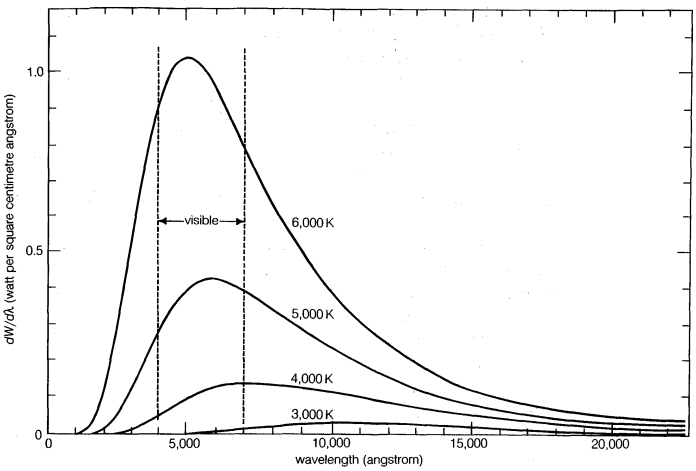


Figure 9: Electromagnetic energy dW emitted per unit area and per second into a wavelength interval, $d\lambda$ = one angstrom, by a blackbody at various temperatures between 3,000 and 6,000 K as a function of wavelength. The range of visible light is between the vertical dashed lines.

From F.A. Jenkins and H.E. White, *Fundamentals of Optics*, copyright © 1976 by McGraw-Hill, Inc.; reproduced with permission

The peak shifts to the visible yellow when the temperature is $T = 6{,}000$ K, like that of the Sun's surface.

It was the shape of Wien's radiative energy distribution as a function of frequency that Planck tried to understand. The decrease of the radiation output at low frequency had already been explained by Lord Rayleigh (John William Strutt) in terms of the decrease, with lowering frequency, in the number of modes of electromagnetic radiation per frequency interval. Rayleigh assumed that all possible frequency modes could radiate with equal probability, following the principle of equipartition of energy. Since the number of frequency modes per frequency interval continues to increase without limit with the square of the frequency, Rayleigh's formula predicted an ever-increasing amount of radiation of higher frequencies instead of the observed maximum and subsequent fall in radiative power. A possible way out of this dilemma was to deny the high-frequency modes an equal chance to radiate. To achieve this, Planck postulated that the radiators or oscillators can only emit electromagnetic radiation in finite amounts of energy of size $E = h\nu$. At a given temperature T, there is then not enough thermal energy available to create and emit many large radiation quanta $h\nu$. More large energy quanta $h\nu$ can be emitted, however, when the temperature is raised. Quantitatively the probability of emitting at temperature T an electromagnetic energy quantum $h\nu$ is

$$\frac{1}{e^{h\nu/kT} - 1},$$

where k is Boltzmann's constant, well known from thermodynamics. With $c = \lambda\nu$, Planck's radiation law then becomes

Planck's radiation law

$$dW = \frac{8\pi ch\lambda^{-5} d\lambda}{e^{hc/\lambda kT} - 1}.$$

This is in superb agreement with Wien's experimental results when the value of h is properly chosen to fit the results. It should be pointed out that Planck's quantization refers to the oscillators of the blackbody or of heated substances. These oscillators of frequency ν are incapable of absorbing or emitting electromagnetic radiation except in energy chunks of size $h\nu$. To explain quantized absorption and emission of radiation, it seemed sufficient to quantize only the energy levels of mechanical systems. Planck did not mean to say that electromagnetic radiation itself is quantized, or as Einstein later put it, "The sale of beer in pint bottles does not imply that beer exists only in indivisible pint portions." The idea that electromagnetic radiation itself is quantized was proposed by Einstein in 1905, as described in the subsequent section.

Photoelectric effect. Hertz discovered the photoelectric effect (1887) quite by accident while generating elec-

tromagnetic waves and observing their propagation. His transmitter and receiver were induction coils with spark gaps. He measured the electromagnetic field strength by the maximum length of the spark of his detector. In order to observe this more accurately, he occasionally enclosed the spark gap of the receiver in a dark case. In doing so, he observed that the spark was always smaller with the case than without it. He concluded correctly that the light from the transmitter spark affected the electrical arcing of the receiver. He used a quartz prism to disperse the light of the transmitter spark and found that the ultraviolet part of the light spectrum was responsible for enhancing the receiver spark. Hertz took this discovery seriously because the only other effect of light on electrical phenomena known at that time was the increase in electrical conductance of the element selenium with light exposure.

A year after Hertz's discovery, it became clear that ultraviolet radiation caused the emission of negatively charged particles from solid surfaces. Thomson's discovery of electrons (1897) and his ensuing measurement of the ratio m/e (the ratio of mass to charge) finally made it possible to identify the negative particles emitted in the photoelectric effect with electrons. This was accomplished in 1899 by J.J. Thomson and independently by Philipp Lenard, one of Hertz's students. Lenard discovered that for a given frequency of ultraviolet radiation the maximum kinetic energy of the emitted electrons depends on the metal used rather than on the intensity of the ultraviolet light. The light intensity increases the number but not the energy of emitted electrons. Moreover, he found that for each metal there is a minimum light frequency that is needed to induce the emission of electrons. Light of a frequency lower than this minimum frequency has no effect regardless of its intensity.

In 1905 Einstein published an article entitled "On a Heuristic Point of View about the Creation and Conversion of Light." Here he deduced that electromagnetic radiation itself consists of "particles" of energy $h\nu$. He arrived at this conclusion by using a simple theoretical argument comparing the change in entropy of an ideal gas caused by an isothermal change in volume with the change in entropy of an equivalent volume change for electromagnetic radiation in accordance with Wien's or Planck's radiation law. This derivation and comparison made no references to substances and oscillators. At the end of this paper, Einstein concluded that if electromagnetic radiation is quantized, the absorption processes are thus quantized too, yielding an elegant explanation of the threshold energies and the intensity dependence of the photoelectric effect. He then predicted that the kinetic energy of the electrons emitted in the photoelectric effect increases with light frequency ν proportional to $h\nu - P$, where P is "the amount of work that the electron must produce on leaving the body." This quantity P, now called work function, depends on the kind of solid used, as discovered by Lenard.

Reaction to Einstein's idea of light quanta

Einstein's path-breaking idea of light quanta was not widely accepted by his peers. Planck himself stated as late as 1913 in his recommendation for admitting Einstein to the Prussian Academy of Sciences "the fact that he [Einstein] may occasionally have missed the mark in his speculations, as, for example, with his hypothesis of light quanta, ought not to be held too much against him, for it is impossible to introduce new ideas, even in the exact sciences, without taking risks." In order to explain a quantized absorption and emission of radiation by matter, it seemed sufficient to quantize the possible energy states in matter. The resistance against quantizing the energies of electromagnetic radiation itself is understandable in view of the incredible success of Maxwell's theory of electromagnetic radiation and the overwhelming evidence of the wave nature of this radiation. Moreover, a formal similarity of two theoretical expressions, in Einstein's 1905 paper, of the entropy of an ideal gas and the entropy of electromagnetic radiation was deemed insufficient evidence for a real correspondence.

Einstein's prediction of the linear increase of the kinetic energy of photoemitted electrons with frequency of light, $h\nu - P$, was verified by Arthur Llewelyn Hughes, Owen Williams Richardson, and Karl Taylor Compton in 1912. In 1916 Robert Andrews Millikan measured both the frequency of the light and the kinetic energy of the electron emitted by the photoelectric effect and obtained a value for Planck's constant h in close agreement with the value that had been arrived at by fitting Planck's radiation law to the blackbody spectrum obtained by Wien.

Compton effect. Convincing evidence of the particle nature of electromagnetic radiation was found in 1922 by the American physicist Arthur Holly Compton. While investigating the scattering of X rays, he observed that such rays lose some of their energy in the scattering process and emerge with slightly decreased frequency. This energy loss increases with the scattering angle, θ, measured from the direction of an unscattered X ray. This so-called Compton effect can be explained, according to classical mechanics, as an elastic collision of two particles comparable to the collision of two billiard balls. In this case, an X-ray photon of energy $h\nu$ and momentum $h\nu/c$ collides with an electron at rest. The recoiling electron was observed and measured by Compton and Alfred W. Simon in a Wilson cloud chamber. If one calculates the result of such an elastic collision using the relativistic formulas for the energy and momentum of the scattered electron, one finds that the wavelength of an X ray after (λ') and before (λ) the scattering event differ by $\lambda' - \lambda = (h/mc)(1 - \cos\theta)$. Here m is the rest mass of the electron and h/mc is called Compton wavelength. It has the value 0.0243 angstrom. The energy $h\nu$ of a photon of this wavelength is equal to the rest mass energy mc^2 of an electron. One might argue that electrons in atoms are not at rest, but their kinetic energy is very small compared to that of energetic X rays and can be disregarded in deriving Compton's equation.

Resonance absorption and recoil. During the mid-1800s the German physicist Gustav Robert Kirchhoff observed that atoms and molecules emit and absorb electromagnetic radiation at characteristic frequencies and that the emission and absorption frequencies are the same for a given substance. Such resonance absorption should, strictly speaking, not occur if one applies the photon picture due to the following argument. Since energy and momentum have to be conserved in the emission process, the atom recoils to the left as the photon is emitted to the right, just as a cannon recoils backward when a shot is fired. Because the recoiling atom carries off some kinetic recoil energy E_R, the emitted photon energy is less than the energy difference of the atomic energy states by the amount E_R. When a photon is absorbed by an atom, the momentum of the photon is likewise transmitted to the atom, thereby giving it a kinetic recoil energy E_R. The absorbing photon must therefore supply not only the energy difference of the atomic energy states but the additional amount E_R as well. Accordingly, resonance absorption should not occur because the emitted photon is missing $2E_R$ to accomplish it.

Nevertheless, ever since Kirchhoff's finding, investigators have observed resonance absorption for electronic transitions in atoms and molecules. This is because for visible light the recoil energy E_R is very small compared with the natural energy uncertainty of atomic emission and absorption processes. The situation is, however, quite different for the emission and absorption of gamma-ray photons by nuclei. The recoil energy E_R is more than 10,000 times as large for gamma-ray photons as for photons of visible light, and the nuclear energy transitions are much more sharply defined because their lifetime can be one million times longer than for electronic energy transitions. The particle nature of photons therefore prevents resonance absorption of gamma-ray photons by free nuclei.

Recoil-free gamma-ray resonance absorption

In 1958 the German physicist Rudolf Ludwig Mössbauer discovered that recoilless gamma-ray resonance absorption is, nevertheless, possible if the emitting as well as the absorbing nuclei are embedded in a solid. In this case, there is a strong probability that the recoil momentum during absorption and emission of the gamma photon is taken up by the whole solid (or more precisely by its entire lattice). This then reduces the recoil energy to nearly zero and thus allows resonance absorption to occur even for gamma rays.

Wave–particle duality. How can electromagnetic radi-

ation behave like a particle in some cases while exhibiting wavelike properties that produce the interference and diffraction phenomena in others? This paradoxical behaviour came to be known as the wave–particle duality. Bohr rejected the idea of light quanta, and he searched for ways to explain the Compton effect and the photoelectric effect by arguing that the momentum and energy conservation laws need to be satisfied only statistically in the time average. In 1923 he stated that the hypothesis of light quanta excludes, in principle, the possibility of a rational definition of the concepts of frequency and wavelength that are essential for explaining interference.

The following year, the conceptual foundations of physics were shaken by the French physicist Louis-Victor de Broglie, who suggested in his doctoral dissertation that the wave–particle duality applies not only to light but to a particle as well. De Broglie proposed that any object has wavelike properties. In particular, he showed that the orbits and energies of the hydrogen atom, as described by Bohr's atomic model, correspond to the condition that the circumference of any orbit precisely matches an integral number of wavelengths λ of the matter waves of electrons. Any particle such as an electron moving with a momentum p has, according to de Broglie, a wavelength $\lambda = h/p$. This idea required a conceptual revolution of mechanics, which led to the wave and quantum mechanics of Erwin Schrödinger, Werner Heisenberg, and Max Born.

Verification of de Broglie's idea of the wavelike behaviour of particles

De Broglie's idea of the wavelike behaviour of particles was quickly verified experimentally. In 1927 Clinton Joseph Davisson and Lester Germer of the United States observed diffraction and hence interference of electron waves by the regular arrangement of atoms in a crystal of nickel. That same year S. Kikuchi of Japan obtained an electron diffraction pattern by shooting electrons with an energy of 68 keV through a thin mica plate and recording the resultant diffraction pattern on a photographic plate. The observed pattern corresponded to electron waves having the wavelength predicted by de Broglie. The diffraction effects of helium atoms were found in 1930, and neutron diffraction has today become an indispensable tool for determining the magnetic and atomic structure of materials.

The interference pattern that results when a radiation front hits two slits in an opaque screen is often cited to explain the conceptual difficulty of the wave–particle duality. Consider an opaque screen with two openings A and B, called double slit, and a photographic plate or a projection screen, as shown in Figure 10. A parallel wave with a wavelength λ passing through the double slit will produce the intensity pattern on the plate or screen as shown at the right of the figure. The intensity is greatest at the centre. It falls to zero at all locations x_0, where the distances to the openings A and B differ by odd-number multiples of a half wavelength, as, for instance, $\lambda/2$, $3\lambda/2$, and $5\lambda/2$. The condition for such destructive interference is the same as for Michelson's interferometer illustrated in Figure 4. Whereas a half-transparent mirror in Figure 4 divides the amplitude of each wave train in half, the division in Figure 10 through openings A and B is spatial. The latter is called division of wave front. Constructive interference or intensity maxima are observed on the screen at all positions whose distances from A and B differ by zero or an integer multiple of λ. This is the wave interpretation of the observed double-slit interference pattern.

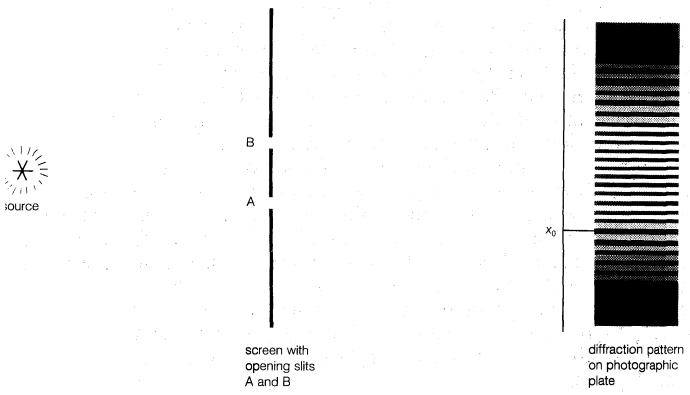


Figure 10: Double-slit interference.

The description of photons is necessarily different because a particle can obviously only pass through opening A or alternatively through opening B. Yet, no interference pattern is observed when either A or B is closed. Both A and B must be open simultaneously. It was thought for a time that one photon passing through A might interfere with another photon passing through B. That possibility was ruled out after the British physicist Geoffrey Taylor demonstrated in 1909 that the same interference pattern can be recorded on a photographic plate even when the light intensity is so feeble that only one photon is present in the apparatus at any one time.

Another attempt to understand the dual nature of electromagnetic radiation was to identify the photon with a wave train whose length is equal to its coherence length $c\tau$, where τ is the coherence time, or the lifetime of an atomic transition from a higher to a lower internal atomic energy state, and c is the light velocity. This is the same as envisioning the photon to be an elongated wave packet, or "needle radiation." Again, the term "photon" had a different meaning for different scientists, and wave nature and quantum structure remained incompatible. It was time to find a theory of electromagnetic radiation that would fuse the wave theory and the particle theory. Such a fusion was accomplished by quantum electrodynamics (QED).

Quantum electrodynamics. Among the most convincing phenomena that demonstrate the quantum nature of light are the following. As the intensity of light is dimmed further and further, one can see individual quanta being registered in light detectors. If the eyes were about 10 times more sensitive, one would perceive the individual light pulses of fainter and fainter light sources as fewer and fewer flashes of equal intensity. Moreover, a movie has been made of the buildup of a two-slit interference pattern by individual photons, such as shown in Figure 10. Photons are particles, but they behave differently from ordinary particles like billiard balls. The rules of their behaviour and their interaction with electrons and other charged particles, as well as the interactions of charged particles with one another, constitute QED.

Subject of QED

Photons are created by perturbances in the motions of electrons and other charged particles; and, in reverse, photons can disappear and thereby create a pair of oppositely charged particles, usually a particle and its antiparticle (*e.g.*, an electron and a positron). A description of this intimate interaction between charged particles and electromagnetic radiation requires a theory that includes both quantum mechanics and special relativity. The foundations of such a theory, known as relativistic quantum mechanics, were laid beginning in 1929 by Paul A.M. Dirac, Heisenberg, and Wolfgang Pauli.

The discussion that follows explains in brief the principal conceptual elements of QED. Further information on the subject can be found in SUBATOMIC PARTICLES: *The development of modern theory;* and MECHANICS: *Quantum mechanics.*

Many phenomena in nature do not depend on the reference scale of scientific measurements. For instance, in electromagnetism the difference in electrical potentials is relevant but not its absolute magnitude. During the 1920s, even before the emergence of quantum mechanics, the German physicist Hermann Weyl discussed the problem of constructing physical theories that are independent of certain reference bases or absolute magnitudes of certain parameters not only locally but everywhere in space. He called this property "Eich Invarianz," which is the conceptual origin of the term "gauge invariance" that plays a crucial role in all modern quantum field theories.

In quantum mechanics all observable quantities are calculated from so-called wave functions, which are complex mathematical functions that include a phase factor. The absolute magnitude of this phase is irrelevant for the observable quantities calculated from these wave functions; hence, the theory describing, for example, the motion of an electron should be the same when the phase of its wave function is changed everywhere in space. This requirement of phase invariance, or gauge invariance, is

Gauge invariance

equivalent to demanding that the total charge is conserved and does not disappear in physical processes or interactions. Experimentally one does indeed observe that charge is conserved in nature.

It turns out that a relativistic quantum theory of charged particles can be made gauge invariant if the interaction is mediated by a massless and chargeless entity which has all the properties of photons. Coulomb's law of the force between charged particles can be derived from this theory, and the photon can be viewed as a "messenger" particle that conveys the electromagnetic force between charged particles of matter. In this theory, Maxwell's equations for electric and magnetic fields are quantized.

The range of a force produced by a particle with nonzero mass is its Compton wavelength h/mc, which for electrons is about 2×10^{-10} centimetre. Since this length is large compared with distances over which stronger nuclear forces act, QED is a very precise theory for electrons.

Despite the conceptual elegance of the QED theory, it proved difficult to calculate the outcome of specific physical situations through its application. Richard P. Feynman and, independently, Julian S. Schwinger and Freeman Dyson of the United States and Tomonaga Shin'ichirō of Japan showed in 1948 that one could calculate the effects of the interactions as a power series in which the coupling constant is called the fine structure constant and has a value close to $1/137$. A serious practical difficulty arose when each term in the series, which had to be summed to obtain the value of an observed quantity, turned out to be infinitely large. In short, the results of the calculations were meaningless. It was eventually found, however, that these divergences could be avoided by introducing "renormalized" couplings and particle masses, an idea conceived by the Dutch physicist Hendrik A. Kramers. Just as a ship moving through water has an enhanced mass due to the fluid that it drags along, so will an electron dragging along and interacting with its own field have a different mass and charge than it would without it. By adding appropriate electromagnetic components to the bare mass and charge—that is, by using renormalized quantities—the disturbing infinities could be removed from the theory. Using this method of renormalization and the perturbation theory, Feynman developed an elegant form for calculating the likelihood of observing processes that are related to the interaction of electromagnetic radiation with matter to any desired degree of accuracy. For example, the passage of an electron or a photon through the double slit illustrated in Figure 10 will, in this QED formalism, produce the observed interference pattern on a photographic plate because of the superposition of all the possible paths these particles can take through the slits.

The success of unifying electricity, magnetism, and light into one theory of electromagnetism and then with the interaction of charged particles into the theory of quantum electrodynamics suggests the possibility of understanding all the forces in nature (gravitational, electromagnetic, weak nuclear, and strong nuclear) as manifestations of a grand unified theory (GUT). The first step in this direction was taken during the 1960s by Abdus Salam, Steven Weinberg, and Sheldon Glashow, who formulated the electroweak theory, which combines the electromagnetic force and the weak nuclear force. This theory predicted that the weak nuclear force is transmitted between particles of matter by three messenger particles designated W^+, W^-, and Z, much in the way that the electromagnetic force is conveyed by photons. The three new particles were discovered in 1983 during experiments at the European Organization for Nuclear Research (CERN), a large accelerator laboratory near Geneva. This triumph for the electroweak theory represented another stepping stone toward a deeper understanding of the forces and interactions that yield the multitude of physical phenomena in the universe. GUT

BIBLIOGRAPHY. Accounts of the historical development of electromagnetic theories may be found in ISAAC ASIMOV, *The History of Physics* (1984); I. BERNARD COHEN, *Revolution in Science* (1985); and THOMAS S. KUHN, *Black-Body Theory and the Quantum Discontinuity, 1894–1912* (1978, reprinted 1987). Early works include EDMUND WHITTAKER, *A History of the Theories of Aether and Electricity,* rev. and enlarged ed., 2 vol. (1951–53); and HEINRICH HERTZ, *Electric Waves: Being Researches on the Propagation of Electric Action with Finite Velocity Through Space* (1893, reissued 1962; originally published in German, 1892). IVAN TOLSTOY, *James Clerk Maxwell* (1981), recounts the life of this pivotal figure, as well as his theory and its ramifications. *James Clerk Maxwell: A Commemoration Volume, 1831–1931* (1931), includes essays by Max Planck and Albert Einstein, among others. Extensive treatments of visible radiation (light) are given by MICHAEL I. SOBEL, *Light* (1987); MAX BORN and EMIL WOLF, *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light,* 6th ed. (1987); and FRANCIS A. JENKINS and HARVEY E. WHITE, *Fundamentals of Optics,* 4th ed. (1976). Classical radiation and electron theory are treated in JOHN DAVID JACKSON, *Classical Electrodynamics,* 2nd ed. (1975); and RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* 3 vol. (1963–65; vol. 1 and 2 have been reprinted, 1977). Wave–particle dualism is addressed by LOUIS DE BROGLIE, *Matter and Light* (1939, reissued 1955; originally published in French, 1937); S. DINER *et al.* (eds.), *The Wave–Particle Dualism* (1984); and A.B. ARONS, *The Development of Concepts of Physics: From the Rationalization of Mechanics to the First Theory of Atomic Structure* (1965). Quantum electrodynamics is discussed in RICHARD P. FEYNMAN, *QED: The Strange Theory of Light and Matter* (1985); RODNEY LOUDON, *The Quantum Theory of Light,* 2nd ed. (1983); W. HEITLER, *The Quantum Theory of Radiation,* 3rd ed. (1964, reprinted 1984); J.M. JAUCH and F. ROHRLICH, *The Theory of Photons and Electrons: The Relativistic Quantum Field Theory of Charged Particles with Spin One-half,* 2nd expanded ed. (1976); and PAUL DAVIES (ed.), *The New Physics* (1989). (Ht.F.)

Electronics

Electronics encompasses an exceptionally broad range of technology having to do with the motion of electrons and its control for useful purposes. The term originally was applied to the study of electron behaviour and movement. It came to be used in its broader sense with advances in knowledge about the fundamental nature of electrons and about the way in which the motion of these particles could be utilized. Today many scientific and technical disciplines—including physics, chemistry, materials science, mathematics, and electrical and electronic engineering—deal with different aspects of electronics.

Research in these fields has resulted in the development of such key devices as transistors, integrated circuits, lasers, and optical fibres. These in turn have made it possible to manufacture a wide array of electronic consumer, industrial, and military products. Such products range from cellular radiotelephone systems and videocassette recorders to high-performance supercomputers and sophisticated weapons systems. By the mid-1980s the electronics industry had become the largest manufacturing industry in the United States. Japan and the industrialized nations of western Europe also had flourishing electronics industries, while various developing countries—including South Korea, Taiwan, Israel, and Yugoslavia—had experienced significant advances as well.

The impact of electronics on modern life has been pervasive. It can be said that the world is in the midst of an electronic revolution at least as significant as the industrial revolution of the 19th century. Evidence of this is apparent everywhere.

Electronics is essential, for example, in telecommunications. Today an increasing volume of information is transmitted in digital form. Digital techniques, in which signals are converted into groups of pulses, allow the intermingling of voice, television, and computer signals into one very rapid series of pulses on a single channel that can be separated at the receiving end and reconstituted into the signals originally sent. Because the digital pulses can be regenerated perfectly after they become attenuated with distance, no noise or other degradation is apparent at the receiving end.

Electronic controls for industrial machines and processes have made possible dramatic improvements in productivity and quality. Computer-aided design tools facilitate the designing of parts that have complex shapes, such as aircraft wings, or intricate structures, such as integrated circuits. The production of designs of this sort is done by computer-controlled machines that receive instructions directly from the design tools.

Access to knowledge has been made far easier by computerized indexes of scientific and technical journals, which are accessible from centralized services over telephone lines. These central databases are being supplemented by new techniques derived from digital audio and video disc technology, which provide locally at low cost access to vast amounts of information in both text and graphic form.

This article reviews the historical development of electronics, highlighting major discoveries and advances. It also describes some key electronic functions and the manner in which various devices carry out these functions.

For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 125, 127, 712, 721, 735, and 738.

The article is divided into the following sections:

History of electronics

EARLY DEVELOPMENTS

Theoretical and experimental studies of electricity during the 18th and 19th centuries led to the development of the first electrical machines and the beginning of the widespread use of electricity. The history of electronics began to evolve separately from that of electricity late in the 19th century with the identification of the electron by the English physicist J.J. Thomson and the measurement of its electric charge by the American physicist Robert A. Millikan in 1909.

At the time of Thomson's work, the American inventor Thomas A. Edison had observed a bluish glow in some of his early light bulbs under certain conditions, and found that a current would flow from one electrode in the lamp to another if the second one (anode) were made positively charged with respect to the first (cathode). Work by Thomson and his students and by the English engineer John Ambrose Fleming revealed that this so-called Edison effect was the result of the emission of electrons from the cathode, the hot filament in the lamp. The motion of the electrons to the anode, a metal plate, constituted an electric current that would not exist if the anode were negatively charged.

This discovery provided impetus for the development of electron tubes, including an improved version of the X-ray tube by the American engineer William D. Coolidge, and

Fleming's thermionic valve

the use of Fleming's thermionic valve (a two-electrode vacuum tube) in radio receivers. The detection of a radio signal, which is a very-high-frequency alternating current, requires that the signal be rectified (*i.e.*, the alternating current must be converted into a direct current by a device that conducts only when the signal has one polarity but not when it has the other). This is precisely what Fleming's valve (patented in 1904) was able to do. Previously, radio signals had been detected by various empirically developed devices such as the "cat whisker" detector, which was composed of a fine wire (the whisker) in delicate contact with the surface of a natural crystal of lead sulfide (galena) or other semiconductor material. These devices were undependable, lacked sufficient sensitivity, and required constant adjustment of the whisker-to-crystal contact to produce the desired result. Yet, these were the forerunners of today's solid-state devices. The fact that crystal rectifiers worked at all encouraged scientists to continue studying them and gradually to obtain the fundamental understanding of the electrical properties of semiconducting materials necessary to permit the invention of the transistor.

In 1906 Lee De Forest, an American engineer, developed a type of vacuum tube that was capable of amplifying radio signals. De Forest added a grid of fine wire between the cathode and anode of the two-electrode thermionic valve constructed by Fleming. The new device, which

De Forest's Audion

De Forest dubbed the Audion, was thus a three-electrode vacuum tube. In operation, the anode in such a vacuum tube is given a positive potential (positively biased) with respect to the cathode, while the grid is negatively biased. A large negative bias on the grid prevents any electrons emitted from the cathode from reaching the anode; however, because the grid is largely open space, it permits some electrons to pass through it and reach the anode at a less negative bias. Small variations in the grid potential can thus control large amounts of anode current.

ERA OF THE ELECTRON TUBE

Impact of the vacuum tube

The first half of the 20th century was the era of the vacuum tube in electronics. This variety of electron tube permitted the development of radio broadcasting, long-distance telephony, television, and the first electronic digital computers. These early electronic computers were, in fact, the largest vacuum-tube systems ever built. Perhaps the best-known representative is the ENIAC (Electronic Numerical Integrator and Calculator; completed in 1946), which was equipped with 17,468 tubes.

The special requirements of the many different applications led to numerous improvements in vacuum tubes, enabling them to handle large amounts of power, operate at very high frequencies, have greater than average reliability, or be made very compact (the size of a thimble).

The cathode-ray tube, originally developed for displaying electrical wave forms on a screen for engineering measurements, evolved into the television picture tube. Such tubes operate by forming the electrons emitted from the cathode into a thin beam that impinges on a fluorescent screen at the end of the tube. The screen emits light that can be viewed from outside the tube. By deflecting the beam, patterns of light are produced on the screen, creating the desired optical images.

Other specialized types of vacuum tubes, developed or refined during World War II for military purposes, are still used today in microwave ovens and as extremely high-frequency transmitters aboard space satellites. Notwithstanding the remarkable success of solid-state devices in most electronic applications, there are certain specialized functions that only vacuum tubes can perform. These usually involve operation at extremes of power or frequency (see below *Electron tubes*). Vacuum tubes continue to be used as display devices for television sets and computer monitors because other means of providing the function are more expensive, though even this situation is changing.

Vacuum tubes are fragile and ultimately wear out in service. Failure occurs in normal usage either by the effects of repeated heating up and cooling down as equipment is switched on and off (thermal fatigue), which ultimately causes a physical fracture in some part of the interior structure of the tube, or by degradation of the properties of the cathode by the residual gases in the tube. Vacuum tubes also take time (from a few seconds to several minutes), to heat up to operating temperature. This is at least an inconvenience and in some cases a serious limitation to their use. These shortcomings motivated scientists at Bell Telephone Laboratories to seek an alternative to the vacuum tube and led to the development of the transistor.

INVENTION OF THE TRANSISTOR: THE SOLID-STATE REVOLUTION

The invention of the transistor in 1947 by John Bardeen, Walter H. Brattain, and William B. Shockley of the Bell research staff provided the first of a series of new devices with remarkable potential for expanding the utility of electronic equipment. Transistors, along with such subsequent developments as integrated circuits (see below), are made of crystalline solid materials called semiconductors, which have electrical properties that can be varied over an extremely wide range by the addition of minuscule quantities of other elements. The electric current in semiconductors is carried by electrons, which have a negative charge, and also by holes, analogous entities that carry a positive charge. The availability of two kinds of charge carriers in semiconductors is a valuable property exploited in many electronic devices made of such materials.

Exploiting the properties of semiconductors

Early transistors were produced using germanium as the semiconductor material, because methods of purifying it to the required degree had been developed during and shortly after World War II. As the electrical properties of semiconductors are extremely sensitive to the presence of the slightest trace of certain other elements, only about 1 part per 1,000,000,000 of such elements can be tolerated in material to be used for making semiconductor devices.

During the late 1950s research on the purification of silicon succeeded in producing material of a quality suitable for semiconductor devices, and new devices made of silicon were manufactured from about 1960. Silicon quickly became the preferred raw material, because it is much more abundant than germanium and thus intrinsically less expensive. In addition, silicon retains its semiconducting properties at higher temperatures than does germanium. For this reason, silicon diodes can be operated at temperatures up to 200° C (392° F), whereas germanium diodes cannot be operated above 85° C. There was one other important property of silicon that was not appreciated at the time but that proved to be crucial to the development of low-cost transistors and integrated circuits: silicon, unlike germanium, forms a tenaciously adhering oxide film with excellent electrical insulating properties when it is heated to high temperatures in the presence of oxygen. This film is utilized extensively as a mask to permit the desired impurities that modify the electrical properties of silicon to be introduced into it during manufacture of semiconductor devices. The mask pattern, formed by a photolithographic process, permits the creation of tiny transistors and other electronic components in the silicon.

Advantages of silicon over germanium

By 1960 vacuum tubes were rapidly being supplanted by transistors because the latter had become less expensive, did not burn out in service, and were much smaller and more reliable. Computers employed hundreds of thousands of transistors each. This fact, together with the need for compact, lightweight electronic missile guidance systems, led to the invention of the integrated circuit (IC) independently by Jack Kilby of Texas Instruments Incorporated in 1958 and by Jean Hoerni and Robert Noyce of Fairchild Semiconductor Corporation in 1959. Kilby is usually credited with having developed the concept of integrating device and circuit elements onto a single silicon chip, while Noyce is given credit for having conceived the method for integrating the separate elements.

Invention of the integrated circuit

Early ICs contained about 10 individual components on a silicon chip three millimetres (0.12 inch) square. By 1970 the number was up to 1,000 on a chip the same size at no increase in cost. Late in the following year the first microprocessor was introduced. The device contained all the arithmetic, logic, and control circuitry required to perform the functions of a computer's central processing unit. This type of large-scale IC was developed by a team at

Intel Corporation, the same company that also introduced the memory integrated circuit in 1971. The stage was now set for the computerization of small electronic equipment.

Until the microprocessor appeared on the scene, computers were essentially discrete pieces of equipment used primarily for data processing and scientific calculations. They ranged in size from minicomputers, comparable in dimensions to a microwave oven, to mainframe systems that took up enough space to fill a large room. The microprocessor enabled computer engineers to develop microcomputers—systems about the size of a lunch box or smaller but with enough computing power to perform many kinds of business, industrial, and scientific tasks. Such systems made it possible to control a host of small instruments or devices (*e.g.*, numerically controlled lathes and one-armed robotic devices for spot welding) by using standard components programmed to do a specific job. The very existence of computer hardware inside such devices is not apparent to the user.

Microprocessors and microcomputers

The large demand for microprocessors generated by these initial applications led to high-volume production and a dramatic reduction in cost. This in turn promoted the use of the devices in many other applications, as, for example, in household appliances and automobiles, for which electronic controls had previously been too expensive to consider. Continued advances in IC technology gave rise to very-large-scale integration (VLSI), which substantially increased the circuit density of microprocessors. This technological advance, coupled with further cost reductions stemming from improved manufacturing methods, made feasible the mass production of personal computers for use in schools and homes, as well as in offices.

Very-large-scale integration

By the mid-1980s inexpensive microprocessors of high-function densities also stimulated computerization of an enormous variety of consumer products. Common examples included programmable microwave ovens and thermostats, clothes washers and dryers, self-tuning television sets and self-focusing cameras, videocassette recorders and video games, telephones and answering machines, musical instruments, watches, and security systems. Microelectronics also came to the fore in business, industry, government, and other sectors. Microprocessor-based equipment proliferated, ranging from automatic teller machines and point-of-sale terminals in retail stores to automated factory assembly systems and office workstations.

By mid-1986 memory integrated circuits with a capacity of 262,144 bits (binary digits) were available. Within a short time, circuits of this kind that had four times that capacity were being produced. By the late 1980s microprocessors capable of handling 16-bit words were common, and 32-bit versions were even available. The larger memories and microprocessors contained more than 3,000,000 transistors on a silicon chip less than two centimetres square. In addition, literally tens of thousands of other kinds of integrated circuits for various applications were available, varying in complexity from a few dozen transistors upward.

DIGITAL ELECTRONICS

Reference was made earlier to digital forms of communication. These arose largely because of the way computers operate—*i.e.*, by using digital representations of numbers. Computers understand only two numbers, 0 and 1, and do all of their arithmetic operations in the binary mode. Many electrical and electronic devices have two states: they are either off or on. A light switch is a familiar example, as are vacuum tubes and transistors. Because computers have been a major application for integrated circuits from their beginning, digital integrated circuits have become commonplace. It has thus become easy to design electronic systems that use digital language to control their functions and to communicate with other systems.

A major advantage in using digital methods is that the correctness of a stream of digital signals can be verified, and, if necessary, errors can be corrected. In contrast, signals that vary in proportion to, say, the sound of an orchestra can be corrupted by "noise," which once present cannot be removed. An example is the sound from a conventional phonograph record, which always contains some extraneous sound from the surface of the recording groove even when the record is new. The noise becomes more pronounced with wear. Contrast this with the sound from a digital compact disc recording. No sound is heard that was not present in the recording studio. The disc and the player contain error-correcting features that remove any incorrect pulses (perhaps arising from dust on the disc) from the information as it is read from the disc.

Error-correcting capability

As electronic systems become more complex, it is essential that errors produced by noise be removed, otherwise, the systems may malfunction. Many electronic systems are required to operate in electrically noisy environments, such as in an automobile. The only practical way to assure immunity from noise is to make such a system operate digitally. That alone may be insufficient, since error-correcting procedures have limits. In principle, it is possible to correct for any arbitrary number of errors, but in practice this may not be possible. The amount of extra information that must be handled to correct for large rates of error reduces the capacity of the system to handle the desired information, and so tradeoffs are necessary.

A consequence of the veritable explosion in the number and kinds of electronic systems has been a sharp growth in the electrical noise level of the environment. Any electrical system generates some noise, and all electronic systems are to some degree susceptible to disturbance from noise. The noise may be conducted along wires connected to the system, or it may be radiated through the air. Care is necessary in the design of systems to limit the amount of noise that is generated and to shield the system properly to protect it from external noise sources.

OPTOELECTRONICS

Many semiconductor materials other than silicon and germanium exist, and they have different useful properties. Compounds formed by the elements from column III of the periodic table, such as aluminum, gallium, and indium, with those from column V, such as phosphorus, arsenic, and antimony, are of particular interest. These so-called III-V compounds are used to make semiconductor devices that emit light efficiently or that operate at exceptionally high frequencies (see below).

Use of III-V compounds

A remarkable characteristic of these compounds is that they can, in effect, be mixed together. One can produce gallium arsenide, or substitute aluminum for some of the gallium, or also substitute phosphorus for some of the arsenic. When this is done, the electrical and optical properties of the material are subtly changed in a continuous fashion in proportion to the amount of aluminum or phosphorus used.

All of these compounds have the same crystal structure. This makes possible the gradation of composition, and thus the properties, of the semiconductor material within one continuous crystalline body. Modern material-processing techniques allow these compositional changes to be controlled accurately on an atomic scale.

These characteristics are exploited in making semiconductor lasers that produce light of any given wavelength within a considerable range. Such lasers are used, for example, in compact digital audio disc players and as light sources for optical fibre communication (see below).

A new direction in electronics curiously does not make use of electrons but instead employs photons (packets of light). By common consent these new approaches are included in electronics because the functions that are performed are, at least for the present, the same as those performed by electronic systems and because these functions usually are embedded in a largely electronic environment. This new direction is called optical electronics, or optoelectronics.

In 1966 it was proposed on theoretical grounds that glass fibres could be made with such high purity that light could travel through them for great distances. Such fibres were produced during the early 1970s. They contain a central core in which the light travels. The outer cladding is made of glass of a different chemical formulation and has a lower optical index of refraction. This difference in refractive index indicates that light travels faster in the cladding than it does in the core. Thus, if the light beam begins to

Development of optical fibres

move from the core into the cladding, its path is bent so as to move it back into the core. The light is constrained within the core even if the fibre is bent into a circle.

The core of early optical fibres was of such a diameter (several micrometres, or about 1/10 the diameter of a human hair) that the various rays of light in the core could travel in slightly different paths, the shortest directly down the axis and other longer paths wandering back and forth across the core. This limited the maximum distance that a pulse of light could travel without becoming unduly spread by the time it arrived at the receiving end of the fibre, with the central ray arriving first and others later. In a digital communications system, successive pulses can overlap one another and be indistinguishable at the receiving end. Such fibres are called multimode fibres, in reference to the various paths (or modes) that the light can follow (see below *Optical fibres*).

During the late 1970s fibres were made with smaller core diameters in which the light was constrained to follow only one path. This occurs if the core has a diameter about the same as the wavelength of the light traveling in it—*i.e.*, about one or two micrometres (0.001 or 0.002 millimetre, or 0.000039 or 0.000078 inch). These single-mode fibres avoid the difficulty described above. By 1985 optical fibres of this kind capable of carrying light signals more than 160 kilometres (100 miles) became available.

Advantages of optical fibres

Optical fibres have several advantages over the copper wires or coaxial cables so widely used in the past. They can carry information at a much higher rate, occupy less space (an important feature in large cities and in buildings), and are quite insensitive to electrical noise. Moreover, it is virtually impossible to make unauthorized connections to them. Costs, initially high, had dropped by 1985 to the point where most new installations of telephone circuits between central telephone offices and longer distances consisted of optical fibres.

Nearly all current installations use a single light signal traveling in one direction within an optical fibre. The light is provided by a solid-state laser and detected at the receiving end by a semiconductor diode. There is no reason that more than one light signal cannot be present at one time in a fibre; up to 10 such signals have been sent down a single fibre in laboratory tests. Each signal is of a slightly different wavelength and can be separated from the others at the receiving end. Signals also have been sent in both directions simultaneously in the laboratory. New terminal equipment can be retrofitted to allow fibres now in service to carry much more information than they were originally intended to do. The cost of long-distance communication can then be significantly reduced and thereby encourage the use of these circuits for more purposes than at present.

A second phase of optoelectronics was being developed during the late 1980s, but the improved system was not expected to be in service for several years. Given the fact that communication signals arrive at a central switching office in optical form, it is attractive to consider switching them from one route to another by optical means rather than electrically, as is done today. The distances between central offices in most cases are substantially less than the distance light can travel within a fibre. Optical switching would make unnecessary the detection and regeneration of the light signals, steps that are presently required. The principles of an optical central-office switch are already understood, though much research is still needed to provide the new optical components and new manufacturing technology required to produce such a switch.

A third direction in optoelectronics builds in part on the foregoing but to a quite different end. A key problem in developing faster computers and faster integrated circuits to use in them is related to the time required for electrical signals to travel over wire interconnections. This is a difficulty both on the integrated circuits themselves as well as between them. Under the best circumstances, electrical signals can travel in a wire at about 90 percent of the speed of light. A more usual rate is 50 percent. Light, fast as it is, travels about 30 centimetres in a billionth of a second. Modern supercomputers operate at speeds of more than 1,000,000,000 operations per second. Thus, if two signals that start simultaneously from different sites are to arrive at their destination simultaneously, the paths they travel must not differ in length by more than a few centimetres.

Two approaches can be envisioned. In one, all of the integrated circuits are placed as close together as possible to minimize the distances that signals must travel. This creates a cooling problem because the integrated circuits generate heat. In the other possible approach, all the paths for signals are made equal to the longest path. This requires the use of much more wire, because most paths are longer than they would otherwise be. All this wire takes space, which means that the integrated circuits have to be placed farther apart than is preferable.

Ultimately, as computers operate even faster, neither approach will work, and a radically new technique must be used. Optical communication between integrated circuits is one possible answer. Light beams do not take up space or interfere with cooling air. If the communication is optical, then the computation might be done optically as well. Optical computation will require a radically different form of integrated circuit. Such integrated circuits can in principle be made of gallium arsenide and related III-V compounds. Some of these integrated circuits may be useful in an optical central-office switch. These matters are presently under serious study in research laboratories.

Optical computers as a possibility

TRENDS IN MICROELECTRONICS

Microprocessors (see above) have been produced in the past as standard products. A growing trend is toward customizing such standard designs for specific applications. Some microprocessors contain an on-board memory that can be programmed at the time of manufacture to provide specific, fixed properties of the device. It is much less expensive to provide unique microprocessors by this tailoring process than to custom-design new microprocessors from scratch.

Another way to provide customized integrated circuits quickly and at modest cost (as compared with a completely new design) is to provide a stock integrated-circuit design having an array of standard functions that can be interconnected in various ways depending on the user's specific needs. Interconnections are formed at the late stages of manufacture, so stock designs can be tailored quickly and inexpensively.

Electrons move much faster in some III-V compound semiconductors than they do in silicon. Microcircuits made from these materials can thus operate at higher speeds and can be used to produce extremely fast computers. These materials are not yet available in as pure a state as silicon and are much more difficult to process: consequently, integrated circuits using III-V materials are not as complex as those made of silicon. Exploitation of these high-speed properties is expected to continue briskly.

SUPERCONDUCTING ELECTRONICS

Numerous metals completely lose their resistance to the flow of electric current at temperatures approaching absolute zero (0 K, or −273° C) and become superconducting. Other equally dramatic changes in electrical properties occur as well. One of these is the Josephson effect, named for the British physicist Brian D. Josephson, who predicted and then discovered the phenomenon in 1962. The Josephson effect governs the passage of current from one superconducting metal to another through a very thin insulating film between them (the Josephson junction), and the effects of small magnetic fields on this current.

Josephson-junction devices

Josephson-junction devices change from one electrical state to another in extraordinarily short times, offering the possibility of producing superconducting microcircuits that operate faster than any other known kind. Serious efforts have been made to construct a computer on this basis, but most of the projects have been either discontinued or sharply cut back due to technical difficulties. Interest in the approach has also waned because of increases in the speed of III-V semiconductor microcircuits.

Josephson junctions have other uses in science. They make extremely sensitive detectors of small magnetic fields, for example. The voltage across a Josephson junction is known on theoretical grounds to be dependent only

on the values of certain basic physical constants. Since these constants are known to great accuracy, Josephson junctions are now used to provide the absolute standard of direct-current voltage.

Other important applications of Josephson junctions have to do with the metrology of very high-speed signals. Measurements of fast phenomena require the use of even faster measurement tools, which Josephson devices provide.

The science of electronics

FUNDAMENTALS

Since electronics is concerned with the control of the motion of electrons, one must keep in mind that electrons, being negatively charged, are attracted to positive charges and repelled by other negative charges. Thus, in a vacuum electrons tend to space themselves apart from one another and form a cloud, subject to the influences of other charges that may be present. An electric current is created by the motion of electrons, whether in a vacuum, in a wire, or in any other electrically conducting medium. In each of these cases, electrons move as a result of their attraction to positive charges or repulsion from negative ones.

An atom consists of a nucleus of protons and neutrons around which electrons, equal in number to the protons in the nucleus, travel in orbits much like those of the planets around the Sun. Because of this equality in the number of positively and negatively charged constituent particles, the atom as a whole is electrically uncharged. When atoms are combined into certain solids called covalent solids (notably the elements of column IV of the periodic table), the valence (or outer) electrons are shared between neighbouring atoms and the atoms thereby become bound together. This occurs not only in elemental solids, wherein all the atoms are of the same kind, but also in chemical compounds (*e.g.*, the III-V compounds).

Variation in conductivity

Different materials vary greatly in their ability to conduct electricity, depending directly on the ease or difficulty of setting electrons free from their atoms. In insulating materials all the outermost electrons of the atoms (the only ones available for electrical conduction in most cases) are tightly bound in the chemical bonds between atoms and are not free to move. In metals there are more valence electrons than are required for bonding, and these excess electrons are freely available for electrical conduction.

Most insulators and metals are crystalline materials, but are composed of a great many very small crystals. The properties of such materials are not much affected by the size of these small crystals. In all crystals, the atoms are positioned in a regularly spaced three-dimensional array.

Semiconducting solids for electronic applications are prepared as single large crystals. In semiconductors the fact that the atoms are in a periodic, three-dimensional array of large size (large, that is, in comparison with an atom) makes the atoms appear nearly invisible to electrons moving within a crystal. The reasons for this behaviour are too complex to explain here, but this property allows electrons to be quite mobile in semiconductors. (For a detailed treatment of the properties of semiconductors, see MATTER: *Semiconductors and insulators.*)

Conduction in semiconductors. In semiconductors such as silicon (which is used as the example here), each constituent atom has four outer electrons, each of which pairs with an electron from one of the four neighbouring atoms to form the interatomic bonds. Completely pure silicon thus has essentially no electrons available at room temperature for electronic conduction, making it a very poor conductor. If an atom from column V of the periodic table, such as phosphorus, is substituted for an atom of silicon, four of its five outer electrons are used for bonding, but the fifth is free to move within the crystal.

If the replacement atom comes from column III of the periodic table, say, boron, it will have only three outer electrons, one too few to complete the four interatomic bonds. The fact that the crystal would be electrically neutral were this bond complete means that if an electron is missing, the vacancy will have a positive charge. A neighbouring electron can move into the vacancy, leaving another vacancy in the electron's former place. This vacancy, with its positive charge, is thus mobile and is called a "hole." Holes in semiconductors move about as readily as electrons do, but because they are positively charged they move in directions opposite to the motion of electrons.

Movement of electrons and holes

Semiconductors whose principal charge carriers are electrons are called *n*-type (*n* standing for negative). If the charge carriers are mainly holes, the material is *p*-type (*p* for positive). The process of substituting elements for the silicon (in this example) is called doping, while the elements are referred to as dopants. The amount of dopant that is required in practical devices is very small, ranging from about 100 dopant atoms per 1,000,000 silicon atoms downward to 1 per 1,000,000,000.

Producing single-crystal silicon

Dopants may be added to the silicon either during the crystal growth process or later. Growth of silicon crystals begins with the preparation of extremely pure polycrystalline silicon having fewer than 1 dopant atom per 10,000,000,000 silicon atoms. This silicon is melted in a quartz-lined furnace. The temperature of the molten silicon is reduced to just above the melting point, and a small bar (the seed) of silicon in single-crystal form is introduced into the surface of the melt. The molten silicon freezes slowly onto the seed with a crystalline structure that is continuous with the structure of the seed. The seed is slowly withdrawn, usually while rotating, under carefully controlled conditions, and it brings with it a cylindrical ingot of silicon that is a single crystal throughout. This ingot may be up to 200 millimetres in diameter and weigh up to 100 kilograms (220 pounds).

After growth, the silicon crystal is ground to a smooth cylindrical shape and sliced into thin wafers approximately 0.6 millimetre thick using diamond tools. The surfaces of the wafers are polished flat by a series of successively finer abrasives until one side has a perfect mirror finish.

The process of fabricating semiconductor devices is a complex series of more than 200 sequential steps, all of which must be done with utmost precision in an environment cleaner than a hospital operating room. The objective is to add the correct dopants in the proper amounts in the right places. The scale of lateral dimensions in integrated circuits ranges down to one micrometre. A high-power semiconductor device for industrial use, on the other hand, may be so large as to require a slice of silicon measuring well over 125 millimetres in diameter (see also below *Integrated circuits*).

Other considerations. The importance of having a thorough, detailed understanding of all the physical effects related to materials, fabrication processes, and device structures cannot be overstated.

The motion of electrons and holes in semiconductors is governed by the theory of quantum mechanics, which was developed during the 1920s and '30s as a much more general theory of the behaviour of all the elementary particles that make up matter. The electrical and optical effects observed in semiconductor materials, their interactions, and the effects of temperature on them are all understood in considerable detail. This understanding not only makes it possible to explain quantitatively what is observed in laboratory experiments but is essential for predicting how new processes and devices work.

The research necessary to develop such a detailed theoretical and experimental body of knowledge was initiated during the late 1940s and has continued in industrial, university, and government laboratories ever since. It is now possible to design new semiconductor devices to perform in a completely predictable fashion by calculating their performance from theory and from their physical configuration with the aid of computers.

The fabrication processes used to make real devices are not as well understood, although much has been learned. Theoretical designs incorporate assumptions that the materials are entirely pure, that dopants exist only in the proper amounts and distributions, and that the dimensions of structures have the intended values. These assumptions are true in practice only to a limited degree. The success of today's semiconductor industry in manufacturing its products is due in large part to the intuition and experience of process engineers, which enable them to cope with

these practical limitations. Large sums of money are spent to provide equipment and manufacturing environments that adequately control each process step and protect the material being processed from contamination.

BASIC ELECTRONIC FUNCTIONS

Rectification. The process of rectification, or conversion of alternating current (AC) to direct current (DC), was briefly mentioned earlier in the discussion of vacuum tubes. A diode, or two-terminal, device is required for this process. Semiconductor diodes consist of a crystal, part of which is *n*-type and part *p*-type. The boundary between the two parts is called a *p–n* junction. As noted above, there is a population of holes on the *p*-type side of the junction and a population of electrons on the *n*-type side.

Use of semiconductor diodes

If a negative voltage is applied to the *p*-type side, implying a positive voltage applied to the *n*-type side, the holes in the *p*-type region will be attracted away from the *p–n* junction, as will the electrons on the *n*-type side. A region on either side of the *p–n* junction will be depleted of charge carriers, thus becoming effectively an insulator. In this condition, called reverse bias, only a very small leakage current flows.

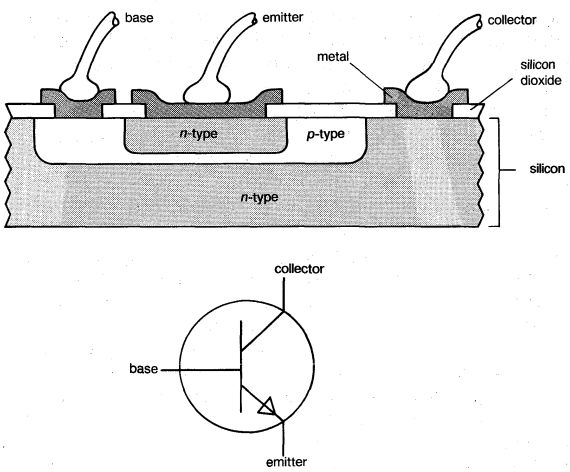


Figure 1: (Top) Cross section of an *n–p–n* transistor and (bottom) its electronic symbol.

If these voltages are reversed (forward bias), the positive voltage on the *p*-type side repels holes across the *p–n* junction; the negative voltage also repels the electrons on the *n*-type side. Both holes and electrons cross the *p–n* junction in opposing directions, creating an electric current.

There are many details of the motion of the holes and electrons that have been omitted in this simple description, but the principle seems clear. The *p–n* junction in a semiconductor diode conducts current with one polarity of applied voltage but not with the other polarity. Typical small diodes will conduct about 0.1 ampere (A) with roughly a 1.5-volt (V) forward bias and withstand 100 or more volts with negligible current flow in the reverse direction. Large industrial diodes can carry up to 5,000 amperes and block several thousand volts.

Amplification. A transistor is constructed with two *p–n* junctions parallel and very close to one another. Figure 1 (top) shows an *n–p–n* transistor with a connection to each of the three regions thus defined. The corresponding electronic symbol for an *n–p–n* transistor is also given in Figure 1 (bottom). The device is not symmetrical but has different levels of doping in the two *n*-type regions and other features that improve its efficiency.

An *n–p–n* transistor

The three regions are labeled as emitter, base, and collector in Figure 1. In normal operation, such as in the amplifier circuit shown in Figure 2, there are provisions (batteries in this case) for applying a small forward bias to the base–emitter junction and a larger reverse bias to the base–collector junction. Resistors are shown in series with each battery to establish the steady-state operating conditions of the circuit. In the base lead is an AC signal source.

If the AC signal source is switched off, the battery in the emitter–base circuit causes a small current to flow

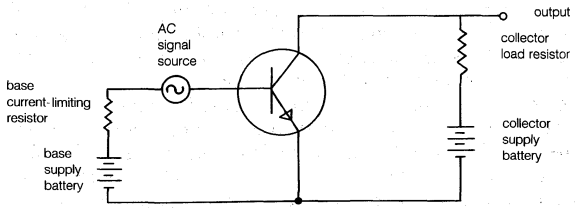


Figure 2: Amplifier using an *n–p–n* transistor.

through the series resistor and the forward-biased emitter–base junction. This results in excess electrons being present in the *p*-type base region of the transistor. Many more of these electrons are attracted to the collector region by the strong reverse bias on the collector than are attracted to the base connection. In an average *n–p–n* transistor more than 100 electrons pass from the emitter to the collector for each one that passes from the emitter to the base.

If the AC signal source is switched on, the base current is increased and decreased alternately. The collector current varies in the same way but to a hundredfold larger extent; in effect, the signal has been amplified. The varying collector current through the collector series resistor causes a varying voltage drop, which may be used as the signal source for a subsequent amplifying circuit. This example employs an *n–p–n* transistor. With a *p–n–p* transistor, the action is similar except that holes are the primary charge carriers, and the voltages of the batteries and thus the direction of current are reversed.

Another important type of transistor that was developed by the early 1960s is the field-effect transistor, illustrated in Figure 3. The physical structure is shown at the top and the electronic symbol at the bottom. The example is a metal-oxide-semiconductor field-effect transistor, or MOSFET. Another type, the junction field-effect transistor, works in a similar fashion but is much less frequently used. The MOSFET consists of two regions: the source (here shown connected to the silicon substrate) and the drain of one conductivity type embedded in a body of the opposite conductivity type. The space between the source and the drain is covered by a thin layer of silicon dioxide formed by heating the silicon in an oxidizing atmosphere. A third part of the device, the gate, is a thin metal layer deposited on the silicon dioxide.

Field-effect transistors

The MOSFET shown in Figure 3 is an *n*-channel type. It is so designated because, when in operation, the application of a positive voltage to the gate with respect to the *p*-type region causes a thin conducting region containing mostly electrons to form in the *p*-type region just beneath the gate. The gate voltage repels holes and attracts electrons from the *p*-type region, in which there are some electrons even though the principal charge carriers are holes.

The thin layer of electron-rich material, the channel, connects the source and drain electrically and permits current to flow between them when the drain is biased positively

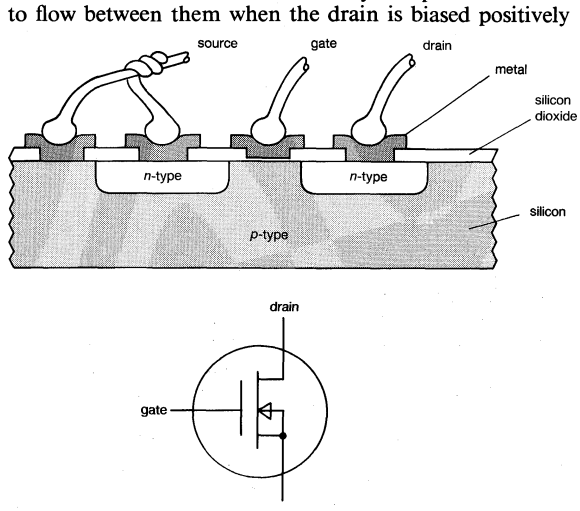


Figure 3: (Top) Cross section of an *n*-channel MOSFET and (bottom) its electronic symbol.

with respect to the source. The amount of current is controlled by the gate voltage. Without gate voltage, no current flows, because the *p–n* junction around the drain region is reverse biased and because no channel exists. MOSFETs are widely used in integrated circuits.

Flexibility in designing solid-state circuits

The existence of more than one type of transistor gives the circuit designer additional freedom not available for vacuum-tube circuits and allows many clever circuits to be constructed. This becomes readily apparent in the direct coupling of successive amplifier stages. There are many ways to couple a signal from one circuit to another. Each has its advantages and disadvantages. Consideration must be given to the voltage levels in the circuits. If both amplifiers were designed as the one in Figure 2, the voltage level at the collector of the first amplifier would be different from that at the base of the second. A direct connection thus could not be used. A transformer could be employed for coupling, with its primary in the collector circuit of the first amplifier and its secondary in the base circuit of the second one. However, transformers often do not exhibit uniform behaviour over a wide range of frequencies, which can be a problem. Transformers also are expensive and bulky. Similarly, a capacitor could be inserted between the collector of the first amplifier and the base of the second. This works well for many applications, providing uniform coupling inexpensively over a wide frequency range. At low frequencies capacitive coupling becomes ineffective, however.

The use of a *p–n–p* second amplifier allows direct connection between the amplifiers, as shown in Figure 4. If properly designed, this arrangement provides useful amplifying properties from direct current to quite high frequencies. Care is required to avoid any changes in the DC operating conditions of the first amplifier, which will cause an amplified change in the DC conditions of the second one. Changes in temperature, in particular, can cause changes in resistor values and changes in the amplification properties of transistors. These factors must be carefully taken into account. Judicious use of feedback from later parts of a circuit to earlier ones can be utilized to stabilize such circuits or to perform various other useful functions (see below). In negative feedback, the feedback signal is of a sense opposite to the signal present at the point in the circuit where the feedback signal is applied. While this has the effect of reducing the overall gain of the circuit, it also corrects numerous small distortions that may have occurred in the signal. For example, if the amplifier does not amplify large signals as much as small ones, the feedback from larger signals will be less, as will the reduction in gain, and the larger signals will be increased in the output of the circuit. Thus the distortion is reduced.

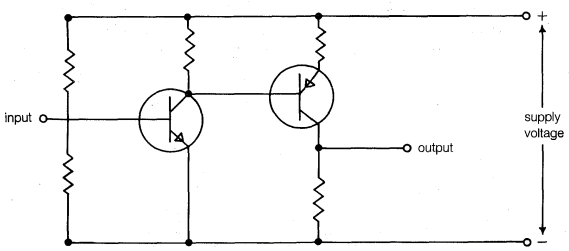


Figure 4: Direct-coupled *n–p–n—p–n–p* amplifier.

Oscillation. If feedback is positive, the feedback signal reinforces the original one, and an amplifier can be made to oscillate, or generate an AC signal. Such signals are needed for many purposes and are created in numerous kinds of oscillator circuits. An example of a tunable oscillator, such as that required for a radio receiver, is shown in Figure 5A. The parallel combination of an inductor and a capacitor is a tuned circuit; at one frequency, and only one, the inductive effects and the capacitive effects balance. At this frequency the voltage developed across the tuned circuit is a maximum. In the oscillator of Figure 5A, positive feedback is provided by the inductor in the collector circuit, which is magnetically coupled to the inductor of the tuned circuit. The connections to these inductors are arranged so that, when the collector current

Tuned circuits

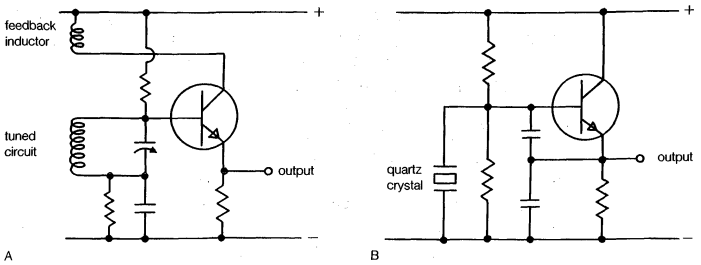


Figure 5: (A) Tunable oscillator; (B) quartz crystal oscillator.

increases, the voltage at the base also increases, thus causing the collector current to rise further. The action of the tuned circuit reverses this sequence after a time and causes the base voltage to start to fall. This reduces the collector current; the positive feedback then further reduces the base voltage, and so on.

The circuit is in fact an amplifier whose output provides the input signal. The tuned circuit affects the feedback process in such a way that the circuit responds to an input signal at only one frequency—namely, the frequency to which the inductor and capacitor are tuned. The variable capacitor provides a way to adjust the frequency of oscillation. The output signal is obtained from the emitter resistor, through which the current rises and falls in synchrony with the collector current.

Oscillators that produce a single, accurate frequency are often needed. Such an oscillator is used in electronic watches. Other circuits in the watch count the output signals from the oscillator to determine the passage of time. These oscillators use a quartz crystal instead of a tuned circuit to establish the operating frequency.

Quartz crystal oscillators

Quartz has the useful property of changing its dimensions slightly if an electric field is applied to it and, conversely, of producing a small electrical voltage when pressure is applied (the piezoelectric effect). In a quartz crystal oscillator a small plate of quartz is provided with metal electrodes on its faces. Just as a bell rings when struck, the quartz plate also "rings," but at a very high frequency, and produces an AC voltage between the electrodes at this mechanically resonant frequency. When used in an oscillator, positive feedback provides energy to the quartz crystal to keep it ringing, and the oscillator output frequency is precisely controlled by the quartz crystal. An oscillator circuit of this kind is shown in Figure 5B.

Quartz is not the only crystalline material that exhibits a piezoelectric effect, but it is used in this application because its oscillation frequency can be rendered to be quite insensitive to temperature changes. Quartz-controlled oscillators are able to produce output frequencies from about 10 kilohertz to more than 200 megahertz and, in carefully controlled environments, can have a precision of one part in 100,000,000,000, though one part in 10,000,000 is more common. The circuit of Figure 5B resembles that of Figure 5A in that the output is taken from the emitter resistor and the tuned circuit is in the base circuit, though the tuned circuit in the crystal oscillator is the quartz crystal. The feedback path is different, however. The feedback signal comes from the emitter and is coupled to the base and to the quartz crystal by the two small capacitors connected in series from the base to the negative supply connection.

Switching and timing. Switching in electronic circuits can be considered in two distinct ways—namely, in terms of digital and non-digital applications.

Non-digital applications. Transistors in amplifier circuits are used as linear devices; *i.e.*, the input signal and the output signal are nearly exact replicas of one another except that the output signal is of a greater amplitude in most cases. Transistors and other semiconductor devices may also be used as switches. In such applications the base or gate of a transistor, depending on the type of transistor in use, is employed as a control element to switch on or off the current between the emitter and collector or the source and drain. The purpose may be as simple as lighting an indicator lamp or it may be of a much more complex nature.

An example of a moderately sophisticated application is in a backup, or "uninterruptible," power source for a computer. Such equipment consists of a storage battery (which is normally kept charged by rectifying the power coming from the AC power line), a circuit for converting the battery power into alternating current, and the necessary control circuits. The control circuits monitor the voltage supplied from the power line. If this voltage varies significantly either upward or downward from its normal values, the control circuit causes the power supply lines to the computer to be switched from the incoming power line to an alternate source of AC derived from the battery.

Inversion by means of transistors

Batteries are usually low-voltage DC sources. Consequently, their energy has to be converted to AC and applied to a transformer, which is useful only in AC circuits, so as to raise the voltage to the proper level for operating the computer. The conversion from DC to AC, a process known as inversion, is often done with high-power transistors operated as switches. The battery is connected to the primary coil of the transformer through the transistors, first in one polarity and then in the other, at a frequency identical to the normal power-line frequency—usually 50 or 60 hertz.

The same result could in principle be obtained by operating the transistors as an oscillator powered by the battery and supplying a smoothly varying AC voltage to the transformer rather than the square pulses obtained via the switching process. This is a much less efficient procedure, however. A transistor operated as a switch is quite efficient, because in its "off" condition very little current flows (a slight leakage through the reverse-biased collector junction) at a relatively high voltage, while in the "on" condition the collector-emitter voltage is very low even though the current is large. In both conditions, the power lost is the product of the voltage and the current. Given this fact, the loss is small because at any instant either the voltage or the current is small.

On the other hand, if the transistor is used as an oscillator, the voltage between the collector and emitter is large most of the time, as is the current, resulting in a substantial loss of power in the transistor as heat. To provide a given amount of power to the computer for a comparable amount of time, a much larger battery and larger and more expensive power transistors would be required, as compared with those needed for a switching mode of operation.

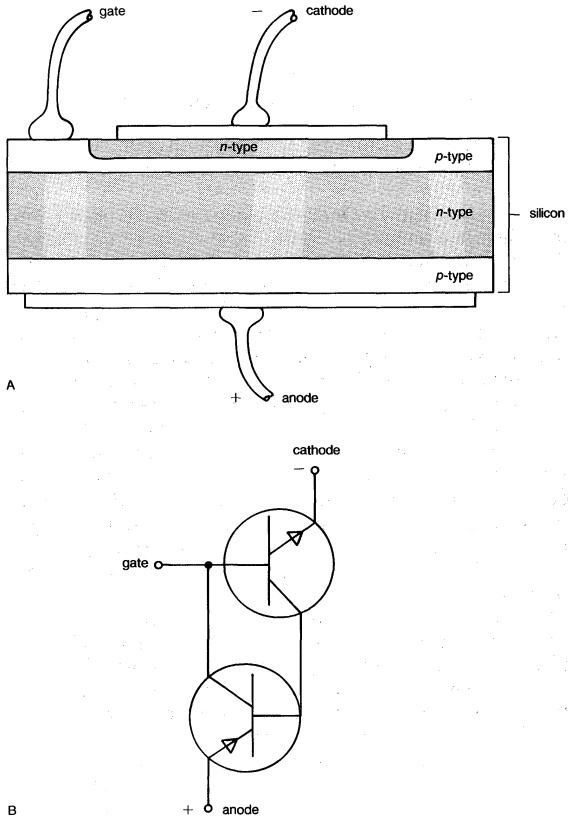


Figure 6: (A) Cross section and (B) *n–p–n—p–n–p* analog of a thyristor.

When the battery is supplying power, the frequency of operation of the power circuits is controlled by a low-power oscillator circuit, which generates timing signals to switch the transistors on at the correct rate. The control circuits monitor the output voltage from the transformer and adjust the "on" period of the transistors by switching them off at the proper time to keep the voltage at the correct level as the power demands of the computer vary.

Use of thyristors in switching applications

Thyristors form another important class of semiconductor devices used in switching applications. Figure 6A illustrates the structure of the simplest of these devices, the controlled rectifier, made of silicon. It may be regarded as two transistors connected, as represented in Figure 6B. If the polarity of the voltage applied between the anode and cathode were the reverse of that shown, the device would not conduct except for minor leakage current across the reverse-biased anode *p–n* junction.

As shown, however, the device will start to conduct if a suitable amount of gate current is applied, but otherwise it will not. The gate current is the equivalent of base current for the *n–p–n* transistor; the resulting larger collector current is the base current for the *p–n–p* transistor. The *p–n–p* transistor has an unusually wide base region, so its gain is small, especially at low currents. Its collector current augments the initial gate current, however. This positive feedback increases the current levels throughout the thyristor, increasing the gain of the *p–n–p* transistor, and at a certain point the combined currents through the *n–p–n* and *p–n–p* transistors are sufficient to maintain conduction through the device even if the gate current is removed. The transistors drive each other into a saturated condition such that the thyristor conducts a large current with a very low voltage drop, typically about one volt. The device remains in this conducting state for an arbitrary period and cannot be turned off under control of the gate. Conduction will cease if the anode polarity becomes negative with respect to the cathode.

Suitability of thyristors for AC circuits

Thyristors are thus well suited for operation in AC rather than DC circuits. They can be switched on during the appropriate half-cycle of voltage (anode positive) and will automatically switch off when the polarity reverses. A single thyristor can be used as a rectifier to produce a variable DC output from a fixed AC input. Adjustment of the DC output is made by modifying the time at which the gate current is applied after the AC voltage crosses zero and becomes the right polarity for conduction. Two thyristors connected in antiparallel (*i.e.*, the anode of each is connected to the cathode of the other) form an AC switch, one thyristor being able to conduct on one half-cycle and the other on the alternate half-cycle. The amount of AC power delivered to the load may be adjusted to any level between zero and full power by appropriate timing of the gate signals to the two thyristors.

Thyristors are designed to handle both small and large amounts of power; the largest ones can withstand up to 5,000 volts in the "off" state and can conduct up to 2,000 amperes in the "on" state. Such a device is contained in an enclosure approximately 150 millimetres in diameter and about 30 millimetres thick and is fitted with either large fins or a water-based cooling arrangement, depending on the application. The power loss in the thyristor in such cases may be as much as four kilowatts, but the total amount of power handled may be up to 1,000 times as large. The efficiency is thus very high.

Other types of thyristors include those in which the gate is able to turn off the thyristor (a useful feature for some applications) and those that can be switched on in either direction of current flow. The latter, called a triac (short for triode alternating current switch), finds wide use in light-duty applications, as, for example, in variable-speed home appliances and light dimmers.

Thyristors have many applications in industrial equipment where substantial amounts of power must be controlled electronically. These applications range from transmission of large amounts of electric power over long

distances, which is more efficient if done as DC rather than AC, to control of heating elements in furnaces and supplying power for electronic equipment. The very large thyristors mentioned earlier are employed in power conversion for DC transmission, both from AC to DC and vice versa.

Other functions. Electronic devices provide a host of different functions too numerous to describe here. Many of these are logical functions that are the basis for the digital integrated circuits discussed later in this article. Others form the interface between integrated circuits and power devices, such as thyristors. When using computers and electronic measuring instruments to control industrial processes, large amounts of power often have to be controlled. The interface circuits are designed to link the "brains" of the computer, for example, with the "muscle" of thyristors.

Interface circuits

Other applications depend on the interactions between light and semiconductor materials mentioned above in connection with optical communication. Such applications include the conversion of sunlight to electricity in solar cells. Most cells of this type consist of silicon diodes in specially designed enclosures to allow sunlight to illuminate them. Silicon is transparent to infrared light; this component of solar radiation passes through a solar cell without generating electricity. The waves of visible light, however, have enough energy to create hole–electron pairs in silicon (the mechanism that results in the absorption of the light). In the vicinity of the *p–n* junction, the holes are attracted toward the electrons on the *n*-type side and the electrons are attracted to the holes on the *p*-type side. This constitutes a current that can be used to power small electric appliances or to charge storage batteries.

There are available special thyristors that use light instead of a gate signal to initiate conduction. They have application in high-voltage systems wherein many thyristors in series must be employed to withstand the voltage. The practical difficulties involved in providing gate signals to all of these thyristors, each at a different electrical potential, are simplified by using optical fibres (which are electrical insulators) to conduct pulses of light to the thyristors. The interaction of the light with the silicon produces carriers just as in a solar cell; these carriers provide the gate signal to switch on the thyristors.

Light-emitting diodes (LEDs) are used in many electronic systems as visual indicators. The first LEDs produced red light, but varieties that emit yellow and green light have become common. These LEDs are made from III-V compounds related to gallium arsenide; the ones that generate red light are usually composed of gallium arsenide phosphide.

Laser diodes

Laser diodes, also made of III-V compounds, are used in digital audio and video disc players to read the minuscule tracks molded into the disc and containing the digitally recorded information. Lasers are employed because laser light can be focused into an extremely tiny spot (substantially smaller than can be obtained with conventional light sources) of great brightness. The light scattered from the markings on the disc is detected by semiconductor photodiodes (see below *Light-emitting diodes and semiconductor lasers*).

Electronic devices must perform their functions over a very wide range of frequencies—from DC to tens of billions of operations per second. Individual types of devices vary greatly in their speed of operation, which is one reason why there are so many types. Thyristors are intrinsically slow, for example, and are primarily used at power-line frequencies. Ordinary small transistors are useful at frequencies up to perhaps 300 megahertz, but power transistors such as those employed in stereo amplifiers are not useful much beyond a few megahertz.

In general, small devices operate better at high frequencies than do large ones, principally because unavoidable parasitic internal series resistances and parallel capacitances that limit high-frequency performance can be better dealt with in small devices. One approach to overcoming the problems of high-frequency, high-power operation involves designing large devices as assemblages of very small ones. This is the approach employed with both metal-oxide-semiconductor (MOS) power transistors and conventional bipolar (*n–p–n* or *p–n–p*) transistors for use at higher frequencies. In effect, hundreds of thousands of identical, small transistors are integrated on a single semiconductor chip and perform as one. The use of III-V compounds, in which electrons move more rapidly than in silicon, can also extend the upper limit of operating frequency. The additional cost of manufacturing these unusual devices generally limits their use to specialized applications, such as in mobile radio transmitters, in which their extra performance level is essential.

In summary, it can be said that present-day electronic devices incorporate exotic materials processed with utmost care and function by the manipulation of electrons and holes via electrical and optical means. The conversion of electronic charge carriers into light and the reverse provide remarkable flexibility for many applications. All of these, taken in combination, have resulted in marvels not even dreamed of only a few years ago. (R.I.S.)

Principal devices and components

ELECTRON TUBES

The term electron tube, as suggested earlier, is the generic name for a class of devices that includes vacuum tubes, phototubes, gas-filled tubes, cathode-ray tubes, and photoelectric tubes. An electron tube typically consists of two or more electrodes enclosed in a glass or metal-ceramic envelope, which is wholly or partially evacuated. Its operation depends on the generation and transfer of electrons through the vacuum from one electrode to another, as shown in Figure 7.

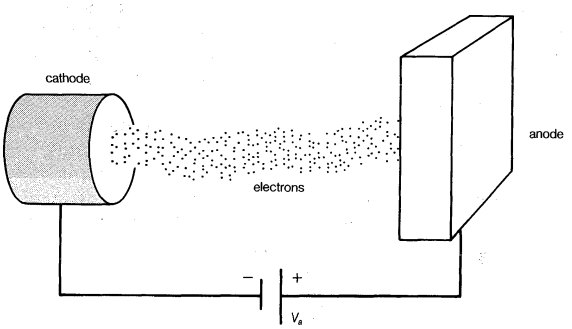


Figure 7: Elements of the simplest electron tube, the diode.

The source of electrons is the cathode, usually a metallic electrode that releases a stream of electrons by one of several mechanisms (*e.g.*, thermionic, secondary, or field emission; see below). Once the electrons have been emitted, their movement is controlled by either an electric field or a magnetic field. An electric field is established by the application of a voltage between the electrodes in the tube (Figure 7), while a magnetic field may be produced external to the tube by an electromagnet or a permanent magnet whose field penetrates the vacuum envelope and influences the emitted electrons. In its simplest form, an electron, being negatively charged, is attracted and accelerated by an electric field from the positive electrode and is repelled and slowed by the field from the negative electrode. The electric fields therefore can be used to change the path of the electron stream, alter the number of electrons flowing (and thereby change the current), and modify their speed. The magnetic field serves primarily to control the movement of the electrons from one electrode to another.

Electron tubes have certain unique properties that cannot be surpassed by solid-state devices for particular applications. Their thermal ruggedness, operating efficiency, and high-power capabilities are features well beyond those provided by the solid-state devices. As components of electronic systems, electron tubes are used as amplifiers, rectifiers, signal generators, and switches. Their applications are extensive, covering a wide range of power levels and frequencies. Electron tubes remain the dominant devices for applications requiring microwave frequencies (above 1,000 megahertz) and moderate- to high-power output.

Electron emission. In its most general sense, electron emission results from directing energy in the form of light, heat, high electric fields, or collisions to a material such that electrons within the material are given enough energy to overcome an energy barrier. Electron emission by application of heat is the mechanism most widely used in electron tubes. The energy barrier is a unique property of a solid and is known as the work function. Thus the ideal materials for use as cathodes in electron tubes are those that yield the lowest work functions. Materials commonly employed for thermionic emission are barium, strontium, and thorium because, when their oxides are combined with base metals of nickel and tungsten, work functions of 1.2 to 3.5 electron volts (eV) are produced. For example, the work functions of tungsten (4.5 eV) and barium (2.3 eV) are reduced significantly through the action of adsorbed atoms on the surface of the emitter where dipoles are formed between the surface and a monolayer of adsorbed atoms (*i.e.*, a single continuous layer of such atoms that is one atom in thickness). In this manner, a tungsten–osmium matrix yields a work function of 1.8 eV when a monolayer of barium oxide is adsorbed on the surface.

Work function

Thermionic emission. When solids are heated to high temperatures, about 1,000° C or higher, electrons can be emitted from the surface. Thermionic emission is not thoroughly understood, but researchers have been able to describe mathematically the emission phenomena using wave mechanics. The most popular models rest on the Richardson–Dushman equation derived in the 1920s and the Langmuir–Child equation formulated shortly thereafter. The former states that the current per unit of area, J, is given by

$$J = AT^2 e^{-W/kT}, \tag{1}$$

where k is Boltzman's constant, A is a constant of the material and its surface finish and is theoretically about 120 amperes per square centimetre per degree Kelvin, T is the temperature of the solid, and W is its work function. When enough electrons are emitted from the cathode of an electron tube, an electron cloud can form in front of the cathode. Such a cloud acts to repel low-energy electrons, and these return to the cathode. This limiting mechanism is aptly referred to as the space-charge-limited operation.

Space-charge-limited operation

In a device such as the diode, the positive voltage applied to the anode attracts electrons from the cloud. The higher the voltage, the more electrons flow to the anode. The higher the voltage, the more electrons flow to the anode. The current density, J, in the space-charge-limited operation, is described by the Langmuir–Child law

$$J = 2.33 \times 10^{-6} \frac{V_a^{3/2}}{d^2}, \tag{2}$$

where V_a is the anode voltage and d is the distance between the anode and the cathode. The key characteristics of thermionic emission, as observed and predicted by equations (1) and (2), are the two regions of emission shown in Figure 8. Much research has been concerned with the transition between the regions and with decreasing the work function of the cathode materials.

Secondary emission. When a metal or dielectric is bombarded by ions or electrons, electrons may be emitted from the surface of the material. The bombarding electrons are called primary and the emitted electrons are designated secondary. Electrons beneath the surface are emitted when they acquire kinetic energy from the primary electrons. The amount of secondary emission depends on the properties of the material, the energy of the primaries, and their angle of incidence. Material properties are characterized by the secondary-emission ratio, which is defined as the number of secondary electrons per primary electron. Typically, the maximum secondary-emission ratio lies between 0.5 and 1.5 for pure metals and occurs for incident electron energies between 200 and 1,000 electron volts. The approximate energy distribution of secondary electrons emitted from a pure metal is skewed such that about 85 percent of the electrons have energies less than 20 electron volts. Positive ion bombardment also can cause secondary emission, but it is much less efficient than electron bombardment. This is because the ions are heavy compared to electrons, and only a small fraction of an ion's energy can be imparted to an electron. Rough or carbon-coated surfaces have relatively small secondary emission.

Secondary-emission ratio

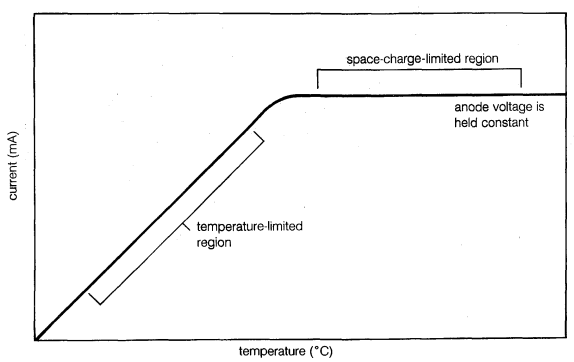


Figure 8: Typical emission characteristic of a thermionic emitter.

Field emission. Electron emission is influenced by an electric field applied at the cathode. For very strong electric fields, the electron emission becomes independent of temperature because the potential barrier at the surface of the cathode is made extremely narrow and electrons tunnel through the barrier even when they have low kinetic energy. Electric fields must be on the order of 10^9 volts per metre in order to cause field emission.

Electron motion in vacuum. Fundamental to all electron devices are the dynamics of charged particles under different electric- and magnetic-field configurations. The charged particles that predominate in the vast majority of applications are electrons (in vacuum). Theoretical treatment of electron trajectories in an evacuated space has made significant strides since the advent of high-speed computers. The fundamental physical laws that serve as the backdrop for the numerical analysis performed by computers are explained below.

The motion of an electron in a uniform field is given by the simple application of Newton's second law, Force = mass × acceleration, in which the force is that exerted on the electron by an applied electric field, $\boldsymbol{E}$. Mathematically, the equation of motion of an electron in a uniform field is given by

$$\boldsymbol{F} = -e\boldsymbol{E} = m\frac{d\upsilon}{dt}, \tag{3}$$

in which e is electron charge 1.60×10^{-19} coulombs, $\boldsymbol{E}$ denotes the field in volts per metre, m is the electron mass 9.107×10^{-31} kilogram, and $d\upsilon/dt$ denotes the rate of change of velocity, which is the electron's acceleration.

When there is a magnetic field present, the electron experiences a second force but only when the electron is in motion. The force is proportional to the product of charge and the velocity component perpendicular to the $\boldsymbol{E}$ field and the magnetic flux density, $\boldsymbol{B}$. The force is directed perpendicular to both the electric field and the electron velocity. Thus an electron traveling at right angles to a uniform magnetic field is deflected in a direction perpendicular to both magnetic and electric fields.

Because the force is always perpendicular to the velocity, the electron trajectory will trace out a perfect circle and maintain that motion at a rate called the cyclotron frequency, ω_c, given by $e/m\boldsymbol{B}$. The circle traced out by the electron has a radius equal to $m\upsilon/e\boldsymbol{B}$. This circular motion is exploited in many electron devices for generating or amplifying radio-frequency (rf) power.

An electron traveling parallel to a uniform magnetic field is unaffected, but any departure from parallelism gives rise to a perpendicular component of velocity and thus a force. This force gives the nearly parallel electron a helical motion about the direction of the magnetic field, keeping it from diverging far from the parallel path. The equation of motion in any of these instances is

$$m\frac{d\upsilon}{dt} = \boldsymbol{B}e\upsilon \sin\theta, \tag{4}$$

where υ is the velocity of the electron in metres per second

in the perpendicular direction to the plane of $\boldsymbol{B}$ and v, and θ is the angle between the directions of $\boldsymbol{B}$ and v. The magnetic flux density is expressed in webers per square centimetre (one weber per centimetre^2 = 10^4 gauss = $10^7/4\pi$ amperes per metre).

Of interest, too, is the situation when the $\boldsymbol{B}$ and the $\boldsymbol{E}$ fields are perpendicular to each other. This configuration is used in beam-focusing devices as well as in a class of devices called magnetrons (see below). In this case, the motion of the electrons is a combination of translation and circular trajectories. The resultant electron trajectory is a cycloid.

Equations 3 and 4 are sufficient to solve for the path and time of transit of electrons in an electron tube except that they require $\boldsymbol{E}$ and $\boldsymbol{B}$ to be known, and these may depend on the presence of electrons or ions. The currents in electron tubes are small enough in most cases that their effect on the magnetic field is usually negligible. The cumulative effect of the electron or ion charge (called space charge) on the electric field cannot always be neglected, however, and this introduces computational difficulty unless the geometry is simple. Furthermore, the electrode currents are so dependent on space charges that the performance characteristics of electron tubes are largely determined by these charges. The electric field with or without space charge can be determined by Gauss's theorem of electrostatics, which states how electric fields are associated with charges. Basically, the rate of change of $\boldsymbol{E}$ with distance is equal to ρ/ε_0 in which ρ is the electric charge density in coulombs per metre, and ε_0 is the permittivity 8.95×10^{-12} farads per metre.

Significance of space charge

The current per unit area, i, entering any surface, as that of an electrode in a tube, is the time rate of change of charge at that surface. This current is the sum of two components, one constituting the actual arrival of electrons at the electrode and the other resulting from the change of induced charge by any change of the electric field with time. Thus, i is the sum of $\rho v + \varepsilon_0 d\boldsymbol{E}/dt$, where v is the electron density and $d\boldsymbol{E}/dt$ is the time-varying electric field. At low frequencies of operation or under steady conditions, the second term is not important. The contrary is true at high frequencies. This equation and the one relating the electric fields to the charges are fundamental to all high-vacuum electron tube phenomena and are sufficient to obtain theoretical solutions.

Energy transfer. The fundamental importance of a large class of electron tubes lies in their ability to amplify power. This power amplification results from the conversion of the energy stored in an external supply via the kinetic energy of electrons to an output energy in the load circuit of the amplifier. The mechanism that makes this conversion possible is electron transport. The underlying principle involved here is that an electron caused to accelerate or decelerate by an electric field will gain or lose kinetic energy. Because energy is conserved, the rf field will increase if the electrons lose kinetic energy, and, conversely, decrease if the electrons gain energy.

The role of electron transport

If a modulated electron convection current flows in an electric field of the same modulation frequency, the power transfer, P, between the field and the electron is

$$P = \frac{1}{2} l_c \boldsymbol{E}, \tag{5}$$

where l_c is the electron convection current and $\boldsymbol{E}$ is the electric field. Both l_c and $\boldsymbol{E}$ are complex quantities, which can be expressed as

$$l_c = l_0(\cos \varphi_l + j \sin \varphi_l) \tag{6}$$

$$\boldsymbol{E} = \boldsymbol{E}_0(\cos \varphi_E + j \sin \varphi_E), \tag{7}$$

in which φ_l and φ_E are the phase angles of the modulated convection current and electric field, respectively. By substituting equations 6 and 7 into 5 and separating the real and imaginary parts, one obtains

$$P_{\text{real}} = \frac{1}{2} l_c E \cos(\varphi_l - \varphi_E) \tag{8}$$

$$P_{\text{imag}} = \frac{1}{2} l_c E \sin(\varphi_l - \varphi_E). \tag{9}$$

Insight into the meaning of equations 8 and 9 may be obtained by considering a physical picture. The negative electron flow (convection current) may be supposed to induce positive charges on the electrodes from which the $\boldsymbol{E}$ field emanates. If the phase is proper, meaning that the induced charges constructively add to the current associated with the modulated E field, the $\boldsymbol{E}$ field grows. Thus, in equations 8 and 9, $P_{\text{real}} = \frac{1}{2} l_c E$, and P_{imag} becomes zero. Conversely, if the phases are 180° apart, P_{real} goes to zero and $P_{\text{imag}} = \frac{1}{2} l_c E$ and power is transferred from the field to the electron current. In practice, different methods are used to produce density modulation in an electron beam (see below).

General applications. The application of vacuum electronics is widespread and involves many types of electron tubes. Another class of electron tubes consists of the tubes employed for rectification and switching. All such vacuum and gas devices make use of a cathode to generate an electron beam and either employ the energy-exchange mechanism or merely illuminate a target, as in the case of a television or X-ray tube. These illuminator tubes are treated in BROADCASTING: *Principles of picture transmission and reception;* RADIATION; and MEASUREMENT AND OBSERVATION. The present discussion focuses on those electron tubes that serve as circuit elements, functioning as rectifiers, microwave rf sources, and amplifiers. Of these, the most important are the latter two types because they constitute most of the electron tubes manufactured since the late 1960s. Within this category the main varieties are magnetrons, klystrons, and traveling-wave tubes. Another notable type is the crossed-field amplifier. Special applications have given impetus to the development of microwave power sources capable of generating tremendous amounts of power (up to billions of watts). These devices, called fast-wave tubes, make up a small but potentially significant element of the electron-tube industry.

Vacuum and gas tubes

Until the late 1950s, electron tubes had been used in virtually every kind of electronic device—computers, radios, transmitters, components of high-fidelity sound systems, etc. After World War II the transistor was perfected and solid-state technology came to be used in all low-frequency (below one gigahertz) and low-power applications. The common conception in the early 1960s was that solid-state technology would rapidly take over microwave applications and thereby render the electron tube obsolete. Such has not been the case, however, and what might be termed a quasi-equilibrium has been established between the solid-state and vacuum-electronic technologies, with each occupying a niche where its technological advantages dominate a particular frequency and power range. The ranges for the various technologies are shown in Figure 9, plotted as domains within the power and frequency spectrum. It will be noted that the higher power levels are dominated by electron tubes, and the lower levels by solid-state devices. High power levels have always been required for transmitters, radar systems, and implements

Use of electron tubes in high-power applications

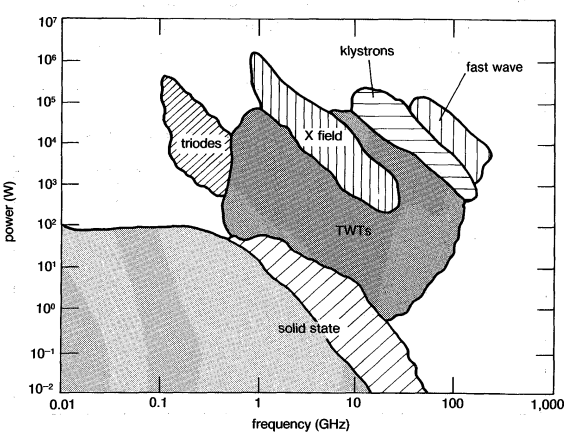


Figure 9: Power and frequency spectrum for sources and amplifiers (see text).

of electronic warfare. Microwave communications systems require power levels of a few watts to hundreds of watts. They include both terrestrial and space microwave links. Extremely high power levels, 10^9 watts, are used primarily for deep-space radars, microwave weapons, and drivers for high-energy particle accelerators.

The prevailing trends in microwave tubes entail improving the amount of bandwidth over which they operate, increasing the efficiency with which DC power is converted to rf power, and enhancing the ability to handle reliably the high power levels. New materials are continually finding use in electron tubes, and this in turn stimulates technological advances. Impressive gains have been scored in cathode technology, in magnets for focusing electron beams in high-thermal conductivity materials, in low secondary-emission materials, and in power-generation techniques.

Electron tubes for rf applications fall into three general categories: (1) space-charge controlled, (2) drift-space, and (3) growing-wave. The space-charge controlled tubes are the triodes, tetrodes, and pentodes. These are low-frequency devices for use below microwave frequencies. A prime example of the drift-space tube is the klystron, which has long been employed in many electronic systems. Two devices, the magnetron and the traveling-wave tube, represent most of the tubes of the growing-wave variety. Some of these and other significant vacuum tubes are delineated below, as are gas tubes employed for rectification and switching.

Klystrons. Devices of this kind are used as amplifiers at microwave frequencies (*e.g.,* in radio relay systems and for dielectric heating) and also as oscillators (*e.g.,* in continuous-wave Doppler radar systems). In klystrons the velocities of electrons emitted from the cathode are modulated to produce a density-modulated electron beam. The principle of operation involved here can be explained in terms of a two-cavity klystron amplifier, the basic structure of which is shown in Figure 10.

Two-cavity klystron amplifier

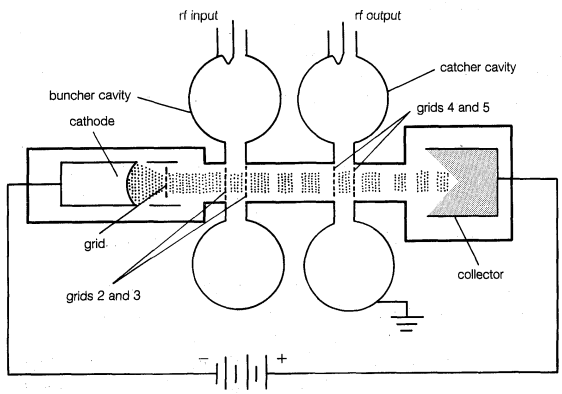


Figure 10: Two-cavity klystron.

The first grid next to the cathode controls the number of electrons in the electron beam and focuses the beam. The voltage between the cathode and the cavity resonators (the buncher and the catcher, which serve as reservoirs of electromagnetic oscillations) is the accelerating potential and is commonly referred to as the beam voltage. This voltage accelerates the DC electron beam to a high velocity before injecting it into the grids of the buncher cavity. The grids of the cavity enable the electrons to pass through but confine the magnetic fields within the cavity. The space between the grids is referred to as the interaction space, or gap. When the electrons traverse this space, they are subjected to rf potentials at a frequency determined by the resonant frequency of the buncher cavity and the input-signal frequency. The amplitude of the rf voltage between the grids is determined by the amplitude of the incoming signal in the case of an amplifier (or by the amplitude of the feedback signal from the second, or catcher, cavity if the klystron is used as an oscillator). Electrons traversing the interaction space when the rf potential on grid 3 is positive with respect to grid 2 are accelerated by the field, while those crossing the gap one half-cycle later are decelerated. In this process, essentially no energy is taken from the buncher cavity, since the average number of electrons slowed down is equal to the average number of electrons speeded up. The decelerated electrons give up energy to the fields inside the cavity, while those that have been accelerated absorb energy from its fields.

Upon leaving the interaction gap, the electrons enter a region called the drift, or bunching, space in which the electrons that were speeded up overtake the slower-moving ones. This causes the electrons to bunch and results in the density modulation of the beam, with the electron bunches representing an rf current in the beam. The catcher is located at a point where the bunching is maximum. This cavity is tuned to the same frequency as the input frequency of the input cavity resonator. The power output at the catcher is obtained by slowing down the electron bunches. If an alternating field exists at the output cavity resonator and grid 4 is positive with respect to grid 5, the electron bunches passing through the grids are decelerated, and they deliver energy to the output cavity. The electron bunches induce an rf current on the walls of the cavity identical to the rf current in the beam. At resonance, the oscillation in the output cavity builds up in proper phase to retard the electron bunches. The power output is equal to the difference in the kinetic energy of the electrons averaged before and after passing the interaction gap. The positive electrode, or collector, located beyond the catcher collects the electrons; it is so designed to minimize secondary emission. (Such emission occurs because of the impact of electrons that reach the end wall.)

Use of the klystron as an oscillator

The klystron amplifier can be converted into an oscillator by employing feedback from the output cavity to the input cavity in proper phase and of sufficient amplitude to overcome the losses in the system. The oscillations are initiated by random fluctuations of the electron beam current and by the noise voltages that generate an alternating field in the cavity. For high-power applications, grids are not used because they would be destroyed by the electron beam.

The output power capability of klystrons is shown in Figure 9. Their power levels are achieved through the use of very high beam voltages and currents. In simple terms the output power is given by $P_o = efficiency \times I_o E_o$, where I_o and E_o are the beam current and voltage and the efficiency is how well the DC power supplied is converted to rf power. For klystrons, the efficiency can be as high as 70 percent. By collecting the spent electron beam at a potential significantly below that of the cavities, higher efficiency can even be achieved—perhaps by as much as 10 to 15 percentage points.

The power gain of the klystron is dependent on the voltage and current as well as on the number of cavities used. The larger the number of cavities employed, the larger the gain that can be obtained. There is, however, a practical limit imposed by the onset of rf instability.

Magnetrons. Electron tubes of this type are primarily used to generate power at microwave frequencies for radar systems and microwave ovens. Magnetrons can be manufactured relatively inexpensively because they require so few parts—namely, a cathode, anode, tank circuit, and magnet. A typical magnetron, shown in Figure 11 in cutaway view, is described below.

The cylindrical anode structure contains a number of equally spaced cavity resonators with slots along the anode surface adjacent to the cylindrical cathode. A permanent

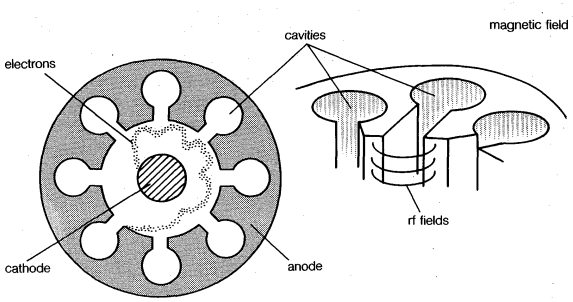


Figure 11: Typical elements of a magnetron.

magnet is usually used to provide the necessary magnetic field. The power output can be coupled through a slot in the cavity or by means of a coupling loop.

As in other types of oscillators, the oscillation originates in random phenomena in the electron space charge and in the cavity resonators. The cavity oscillations produce electric fields that spread outward into the interaction space from the slots in the anode structure, as shown in Figure 11. Energy is transferred from the radial DC field to the rf field by interaction of the electrons with the fringing rf field. The first orbit of an electron in Figure 11 occurs when the rf field across the gap is in a direction to retard the electron. The transfer of energy is from the electron to the tangential component of the rf field. The electron comes to a stop and is again accelerated by the radial DC field, making an orbit adjacent to the next cavity slot. If the rf field across the next cavity slot has changed phase by 180 degrees, the direction of the rf field is in the same direction as that of the electron, which moves in synchrony with the rf field. The electron gives up most of its energy to the cavities before it finally terminates on the anode surface. There is a net delivery of energy to the cavity resonators because electrons that absorb energy from the rf field are quickly returned to the cathode. By contrast, the energy in the rotational component of motion of the electrons in the retarding rf field remains practically unaffected, and the electrons may orbit around the cathode many times.

The phase relationship of the rf fields appearing across adjacent slots is used as a means of defining different modes of oscillation. The mechanism by which electron bunches are formed and by which electrons are kept in synchrony with the rf field is called phase focusing.

Phase focusing

Magnetrons have a wide range of output powers—from that used for cooking, which is about 600 watts, to special ones capable of generating pulsed power levels up to 10^7 watts. The DC-to-rf power-conversion efficiency typically ranges from 50 to 85 percent.

Crossed-field amplifiers. The crossed-field amplifier (CFA) shares several characteristics with the magnetron. Both contain a cylindrical cathode coaxial with an rf structure. Also, each of these tubes constitutes a diode in which a magnetic field is established perpendicular to an electric field between the cathode and anode. Still another similarity between the CFA and magnetron is that the rf structure serves as the electron collector and must therefore be very rugged. The key difference between the CFA and the magnetron is in the rf structure. The CFA uses a delay line that slows down the rf, thereby allowing it to interact more efficiently with the electron stream. With this scheme, CFAs are capable of achieving very high conversion efficiencies of up to 70 percent. Additionally, the output power of CFAs is obtained with relatively low beam voltage, two to three times lower than other devices at the same power level. The gain characteristic of CFAs is a highly nonlinear one and relatively low (one to two orders of magnitude lower) compared with other electron tubes. Bandwidths of CFAs are typically 10 to 20 percent.

Traveling-wave tubes. These are generally used to amplify microwave signals over broad bandwidths. The main elements of a traveling-wave tube (TWT) are (1) an electron gun, (2) a focusing structure that keeps the electrons in a linear path, (3) an rf circuit that causes rf fields to interact with the electron beam, and (4) a collector with which to collect the electrons. There are two main types of TWTs, and these are differentiated by the rf structure. One uses a slow-wave circuit called a helix for propagating the rf wave for electron–rf field interaction, and the other employs a series of staggered cavities coupled to each other for wave propagation. Each type has different characteristics and finds its use in different applications. The helix TWT is distinct from other electron tubes as it is the only one that does not use rf cavities. Because cavities have bandwidth limitations, the coupled-cavity TWT also is bandwidth-limited to typically 10 to 15 percent. The helix TWT, however, has no particular bandwidth limitations and, for all practical purposes, an octave bandwidth (100 percent) is attainable.

Helix and coupled-cavity TWTs

The basic helix TWT is shown schematically in Figure 12. The electron gun contains a cathode that emits electrons, and these are formed by the gun electrodes into a beam that is injected into the opening of the helix.

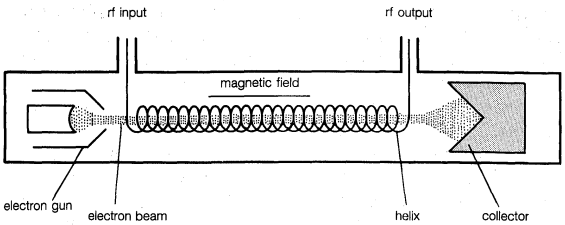


Figure 12: Elements of the traveling-wave tube (TWT).

Because space-charge forces tend to make the electrons diverge radially, a focusing structure is used to keep the beam at a desired diameter by causing diverging electrons to be sent toward the axis of the helix. In this manner, the electron beam is maintained at the desired diameter all along the length of the helix. This is necessary because the electron–rf field interaction takes place continuously over the length of the helix within the helix diameter. In order to achieve this interaction, the diameter and pitch of the helix must be such that the rf wave traveling on the helix wire at the speed of light (*i.e.*, 2.997925×10^8 metres per second) is slowed down in its axial travel to be in synchrony with the velocity of the electrons in the beam. The axial phase velocity of the wave is approximated by multiplying the speed of light by the ratio of the pitch to the circumference of the helix. The axial phase velocity is relatively constant over a wide range of frequencies, and this characteristic provides for the large bandwidths of helix TWTs. For typical applications, the electrons travel down the helix axis at about $1/10$ the speed of light. The voltage required to impart this velocity to the electrons is on the order of 10,000 volts. The rf output power and frequency required determine the actual voltage and current to be used.

The amplifying action of the TWT occurs via a continuous interaction between the axial component of the electric field wave traveling down the centre of the helix and the electron beam moving along the axis of the helix at the same time. The electrons are continually slowed down, and their energy is transferred to the wave along the helix. The electrons tend to bunch in regions where the rf field ahead is decelerating and the field behind is accelerating. The interaction between a bunched electron beam and a helix may be viewed in terms of induced currents. The bunches of electrons induce positive charges on the helix, and these charges move in phase with the bunches. If the phase is proper, this current adds to the current associated with the rf wave flowing in the helix and causes the wave to grow. The interaction is continuous along the length of the helix. The wave amplitude growing on the helix, in turn, causes the electrons to bunch more, and the growing bunches of electrons result in a continuous exponential growth of the helix wave with distance. Typical gains are on the order of four decibels per centimetre, and overall gains are 40 to 50 decibels for helix tubes of practical sizes and applications. The DC-to-rf conversion efficiency of TWTs, both helix and coupled-cavity, is similar and is in the range of 20 to 60 percent, depending on the power level and bandwidth.

A special application of helix TWTs is its use in communications or scientific satellites and other spacecraft. The helix is ideal for this application because of its small size and weight, high efficiency, and low rf-distortion characteristics. TWTs in space have demonstrated ultrareliable operation, amassing millions of hours of operation.

Suitability for use in space vehicles

Fast-wave electron tubes. Conventional electron tubes are designed to produce electron–field interaction by slowing down the rf wave to about $1/10$ the speed of light. A different way of creating the electron–field interaction is to allow the rf wave to propagate at essentially the speed of light by letting it pass, for example, through a section of a wave guide. Electrons used for energy transfer to the fast rf wave are bunched either by rippled magnetic fields or by rf fields that induce angular-velocity modulation.

The bunched electrons give part of their energy to a properly phased microwave field. The advantage of fast-wave devices is that the rf circuits are large compared to the wavelength of a signal. Thus, such devices can be manufactured with large dimensions and still operate at exceedingly high frequencies—*e.g.*, 100 gigahertz or higher. The fast-wave tubes typically operate at very high voltages to generate the high electron velocities required for resonance conditions, thereby permitting an energy exchange to take place. In fact, it is the resonance due to the electrons in a magnetic field that determines the frequency and not a cavity structure, as in a klystron. The high-voltage AC currents used are the main reason that fast-wave devices produce exceedingly high rf power levels, up to millions of watts at very high frequencies (more than 100 gigahertz).

Gyrotrons

One major type of fast-wave electron tube is the gyrotron. Sometimes called the cyclotron resonance maser, this device can generate megawatts of pulsed rf power at millimetre and submillimetre wavelengths. Gyrotrons and other fast-wave tubes are used in certain radar applications, communications, and plasma heating in some experimental controlled-fusion systems.

The gyrotron makes use of an energy-transfer mechanism between an electron orbiting in a magnetic field and an electromagnetic field at the cyclotron frequency (see above; Figure 13). At a very high electron velocity, the electron increases in mass (due to relativistic effects), and this increase lowers the cyclotron frequency. The interaction between the orbiting electron and the electromagnetic field is such that, if energy is given to the field, the electron loses some mass and the phase of the cyclotron wave changes. This results in a form of electron bunching analogous to the bunching in a klystron (see above).

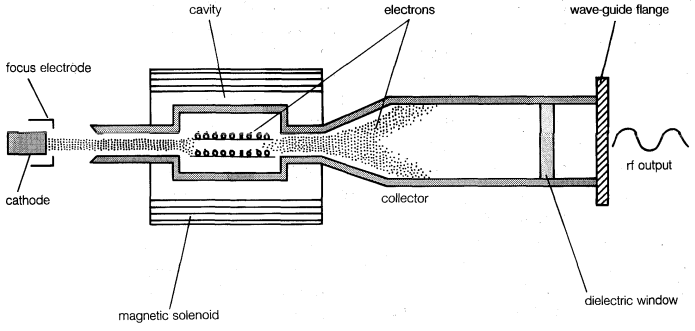


Figure 13: In a gyrotron, electrons from the cathode acquire cyclotron rotation in the cavity (see text).

In another major type of fast-wave tube, an electromagnetic wave travels down a circular or rectangular wave guide and interacts with an undulating electron beam. The undulating motion of the electron beam is produced by a periodic magnetic field. The electrons bunch up as in the klystron process. When the bunches interact with the traveling wave, the electron energy is converted to rf energy, resulting in amplification. Beam voltages in these devices are on the order of 100 kilovolts; and, with electron currents of about 35 amperes, one can generate steady-state power levels of 300 watts or pulsed peak power levels of 200 kilowatts at millimetre wavelengths.

Gas electron tubes. In gas tubes the conductivity between the electrodes differs from that of a vacuum due to the presence of a small amount of gas. Common uses of such devices are rectification and switching (*e.g.*, opening inductive energy-storage circuits, on–off modulations, and closing applications).

Characteristic configuration

The modern gas tube is typically a coaxial, four-electrode device that contains hydrogen gas at a pressure of 50–400 millitorrs (0.000066–0.00053 atmosphere). A low-voltage discharge is initiated near the cathode by the electrons that it generates, and the hydrogen gas molecules are ionized by collisions with the electrons. The electrons released by the ionized hydrogen bombard the cathode, giving rise to secondary electrons. This secondary electron emission sustains the low-voltage discharge. Some primary and secondary electrons are accelerated from the cathode and undergo more collisions with the hydrogen gas molecules. The plasma formed near the cathode can be enlarged so that contact is made with the electrode serving as the anode, and the conduction plasma path is established. The resulting current can be interrupted by means of a control grid with small apertures that pinch off the flow of plasma. (E.N.So.)

SEMICONDUCTOR DEVICES

Semiconductor devices are electronic components fabricated from materials whose electrical conductivity is intermediate between that of an insulator and that of a conductor (see below). They have found wide applications because of their compactness, reliability, and low cost. As discrete components, they can be used to accommodate much of the electromagnetic spectrum—from direct current to ultraviolet frequencies. They have a wide range of current- and voltage-handling capabilities, with current ratings from a few nanoamperes (10^{-9} ampere) to more than 5,000 amperes and voltage ratings extending above 100,000 volts. In addition, semiconductor devices lend themselves to integration into complex but readily manufacturable microelectronic circuits. They are, and will be in the foreseeable future, the key elements for the majority of electronic systems, including communications, consumer, data-processing, and industrial control equipment.

Semiconductor and junction principles. *Semiconductor materials.* Solid-state materials are commonly grouped into three classes: insulators, semiconductors, and conductors. (At low temperatures some conductors, semiconductors, and insulators may become superconductors.) Figure 14 shows the conductivities σ (and the corresponding resistivities $\rho = 1/\sigma$) that are associated with some important materials in each of the three classes. Insulators, such as fused quartz and glass, have very low conductivities on the order of 10^{-18} to 10^{-10} siemens per centimetre; and conductors, such as aluminum and silver, have high conductivities typically from 10^4 to 10^6 siemens per centimetre.The conductivities of semiconductors are between these extremes.

The conductivity of a semiconductor is generally sensitive to temperature, illumination, magnetic fields, and minute amounts of impurity atoms. For example, the addition of less than 0.01 percent of a particular type of impurity can increase the electrical conductivity of a semiconductor by five or more orders of magnitude (*i.e.*, 100,000 times). The ranges of semiconductor conductivity due to impurity atoms for five common semiconductors are given in Figure 14.

Types of semiconductor materials

The study of semiconductor materials began in the early 19th century. Over the years, many semiconductors have been investigated. The Table shows a portion of the periodic table related to semiconductors. The elemental semiconductors are those composed of single species of atoms, such as silicon (Si), germanium (Ge), and gray tin (Sn) in column IV and selenium (Se) and tellurium (Te) in column VI. There are, however, numerous compound semiconductors that are composed of two or more elements. They are called binary compounds (for two elements), ternary compounds (for three elements), and so

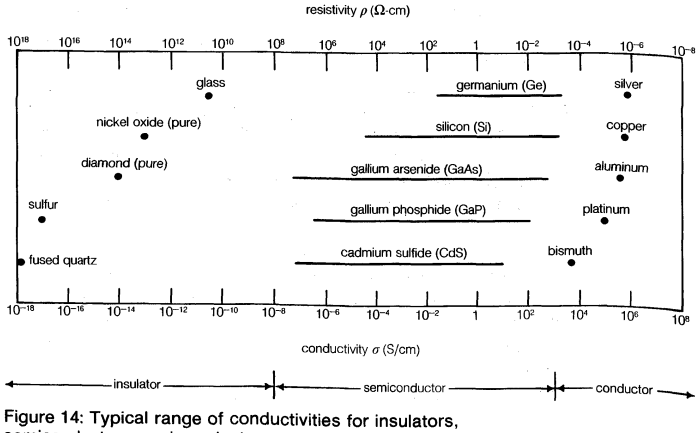


Figure 14: Typical range of conductivities for insulators, semiconductors, and conductors.

Portion of the Periodic Table of Elements Related to Semiconductors

	column				
period	II	III	IV	V	VI
2		boron B	carbon C	nitrogen N	
3	magnesium Mg	aluminum Al	silicon Si	phosphorus P	sulfur S
4	zinc Zn	gallium Ga	germanium Ge	arsenic As	selenium Se
5	cadmium Cd	indium In	tin Sn	antimony Sb	tellurium Te
6	mercury Hg		lead Pb		

forth. Gallium arsenide (GaAs), for example, is a binary III-V compound, which is a combination of gallium (Ga) from column III and arsenic (As) from column V.

Ternary compounds can be formed by elements from three different columns, as, for instance, mercury indium telluride ($HgIn_2Te_4$), a II-III-VI compound. They also can be formed by elements from two columns, such as aluminum gallium arsenide ($Al_xGa_{1-x}As$), which is a ternary III-V compound, where both Al and Ga are from column III and the subscript x is related to the composition of the two elements from 100 percent Al ($x = 1$) to 100 percent Ga ($x = 0$).

Prior to the invention of the bipolar transistor in 1947, semiconductors were used only as two-terminal devices, such as rectifiers and photodiodes. During the early 1950s germanium was the major semiconductor material. However, it proved unsuitable for many applications because devices made of the material exhibited high leakage currents (see below) at only moderately elevated temperatures. Since the early 1960s silicon has become a practical substitute, virtually supplanting germanium as a material for semiconductor fabrication. The main reasons for this are twofold: (1) silicon devices exhibit much lower leakage currents, and (2) high-quality silicon dioxide (SiO_2), which is an insulator, is easy to produce. At present, silicon technology is by far the most advanced among all semiconductor technologies, and silicon-based devices constitute more than 95 percent of all semiconductor hardware sold worldwide.

Many of the compound semiconductors have electrical and optical properties that are absent in silicon. These semiconductors, especially gallium arsenide, are used mainly for high-speed and optoelectronic applications (see below *Optoelectronic devices*).

Electronic properties. The semiconductor materials treated here are single crystals—*i.e.,* the atoms are arranged in a three-dimensional periodic fashion. Figure 15A shows a simplified two-dimensional representation of an intrinsic silicon crystal that is very pure and contains a negligibly small amount of impurities. Each silicon atom in the crystal is surrounded by four of its nearest neighbours. Each atom has four electrons in its outer orbit and shares these electrons with its four neighbours. This sharing of electrons is known as covalent bonding, and each electron pair constitutes a covalent bond. The force of attraction for the electrons by both nuclei holds the two atoms together.

At low temperatures the electrons are bound in their respective positions in the crystal; consequently, they are not available for electrical conduction. At higher temperatures thermal vibration may break some of the covalent bonds. The breaking of a bond yields a free electron that can participate in current conduction. Once an electron moves away from a covalent bond, there is an electron deficiency in that bond. This deficiency may be filled by one of the neighbouring electrons, which results in a shift of the deficiency location from one site to another. This deficiency may thus be regarded as a particle similar to an electron. This fictitious particle, dubbed a hole, carries a positive charge and moves, under the influence of an applied electric field, in a direction opposite to that of an electron. The concept of a hole is analogous to that of a bubble in a liquid. Although it is actually the liquid that moves, it is much easier to talk about the motion of the bubble in the opposite direction.

For an isolated atom, the electrons of the atom can have only discrete energy levels. When a large number of atoms are brought together to form a crystal, the interaction between the atoms causes the discrete energy levels to spread out into energy bands. When there is no thermal vibration, the electrons in a semiconductor will completely fill a number of energy bands, leaving the rest of the energy bands empty. The highest filled band is called the valence band. The next higher band is the conduction band, which is separated from the valence band by an energy gap. This energy gap, also called a bandgap, is a region that designates energies that the electrons in the semiconductor cannot possess. Most of the important semiconductors have bandgaps in the range 0.25 to 2.5 eV. The bandgap of silicon, for example, is 1.12 eV and that of gallium arsenide is 1.42 eV.

Bandgaps of semiconductors

As discussed above, at finite temperatures thermal vibrations will break some bonds. When a bond is broken, a free electron, along with a free hole, results. Because of the relatively small bandgaps of semiconductors, some electrons are able to move from the valence band to the conduction band, leaving holes in the former. When an electric field is applied to the semiconductor, both the electrons in the conduction band and the holes in the valence band gain kinetic energy and conduct electricity.

The electrical conductivity of a material depends on the number of charge carriers (*i.e.,* free electrons and free holes) per unit volume and on the rate at which these carriers move under the influence of an electric field. In an intrinsic semiconductor there exists an equal number of free electrons and free holes. The electrons and holes, however, have different mobilities—that is to say, they move with different velocities in an electric field. For example, for intrinsic silicon at room temperature, the electron mobility is 1,500 square centimetres per volt second ($cm^2/V \cdot s$; *i.e.,* an electron will move at a velocity of 1,500 centimetres per second under an electric field of one volt per centimetre), while the hole mobility is 500 $cm^2/V \cdot s$. The mobilities of a given semiconductor generally decrease with increasing temperature or with increased impurity concentration.

Electrical conduction in intrinsic semiconductors is quite poor. To produce higher conduction, one can intentionally introduce impurities (typically to a concentration of one part per million host atoms). This is the so-called doping process. For example, when a silicon atom is replaced by an atom with five outer electrons such as arsenic, Figure

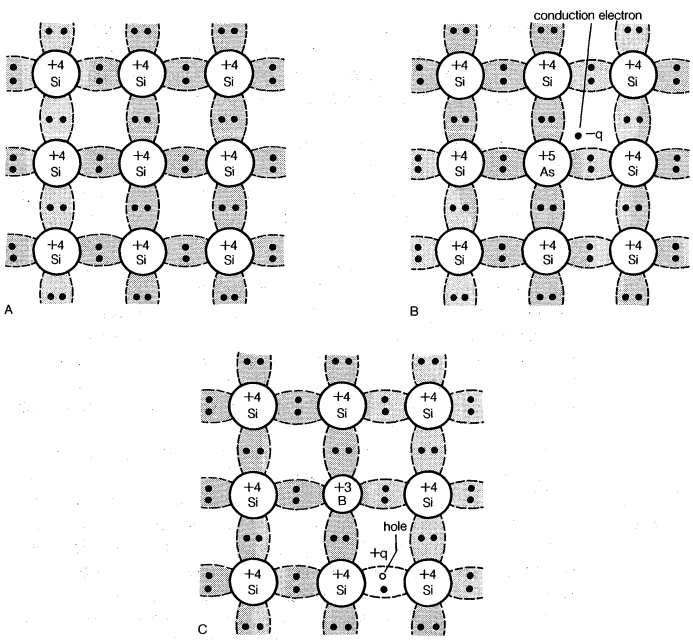


Figure 15: *Three bond pictures of a semiconductor.*
(A) Intrinsic with negligible impurities; (B) *n*-type with donor (arsenic); (C) *p*-type with acceptor (boron).

15B, the arsenic atom forms covalent bonds with its four neighbouring silicon atoms. The fifth electron becomes a conduction electron that is "donated" to the conduction band. The silicon becomes an *n*-type semiconductor because of the addition of the electron. The arsenic atom is the donor. Similarly, Figure 15C shows that, when an atom with three outer electrons such as boron is substituted for a silicon atom, an additional electron is "accepted" to form four covalent bonds around the boron atom, and a positively charged hole is created in the valence band. This is a *p*-type semiconductor, with the boron constituting an acceptor.

The p–n junction. If an abrupt change in impurity type from acceptors (*p*-type) to donors (*n*-type) occurs within a single crystal structure, a *p–n* junction is formed (see the insets of Figure 16). On the *p* side, the holes constitute the dominant carriers and so are called majority carriers. A few thermally generated electrons will also exist in the *p* side; these are termed minority carriers. On the *n* side the electrons are the majority carriers, while the holes are the minority carriers. Near the junction is a region having no free-charge carriers. This region, called the depletion layer, behaves as an insulator.

Majority and minority carriers

The most important characteristic of *p–n* junctions is that they rectify; that is to say, they allow current to flow easily in only one direction. Figure 16 shows the current-voltage characteristics of a typical silicon *p–n* junction. When a forward bias is applied to the *p–n* junction (*i.e.*, a positive voltage applied to the *p*-side with respect to the *n*-side, as shown in the second quadrant), the majority charge carriers move across the junction so that a large current can flow. However, when a reverse bias is applied (in the third quadrant), the charge carriers introduced by the impurities move in opposite directions away from the junction and only a small leakage current flows initially. As the reverse bias is increased, the current remains very small until a critical voltage is reached, at which point the current suddenly increases. This sudden increase in current is referred to as the junction breakdown, usually a nondestructive phenomenon if the resulting power dissipation is limited to a safe value. The applied forward voltage is usually less than one volt, but the reverse critical voltage, called the breakdown voltage, can vary from less than one volt to many thousands of volts, depending on the impurity concentration of the junction and other device parameters.

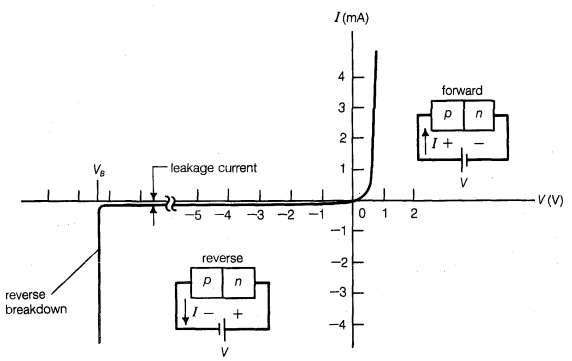


Figure 16: Current-voltage characteristics of a typical silicon *p–n* junction. Insets show forward-bias and reverse-bias conditions.

Two-terminal junction devices. A *p–n* junction diode is a solid-state device that has two terminals. Depending on impurity distribution, device geometry, and biasing condition, a junction diode can perform various functions. There are more than 50,000 types of diodes with voltage ratings from less than one volt to more than 2,000 volts and current ratings from less than one milliampere to more than 5,000 amperes. A *p–n* junction also can generate and detect light and convert optical radiation into electrical energy (see below *Optoelectronic devices*). The circuit symbols of various junction diodes and related semiconductor devices are given in Figure 17.

Junction diodes

Rectifier. This type of *p–n* junction diode is specifically designed to rectify an alternating current—*i.e.*, to give a low resistance to current flow in one direction and a very high resistance in the other direction. Such diodes are generally designed for use as power-rectifying devices that operate at frequencies from 50 hertz to 50 kilohertz. The majority of rectifiers have power-dissipation capabilities from 0.1 to 10 watts and a reverse breakdown voltage from 50 to more than 5,000 volts. (A high-voltage rectifier is made from two or more *p–n* junctions connected in series.)

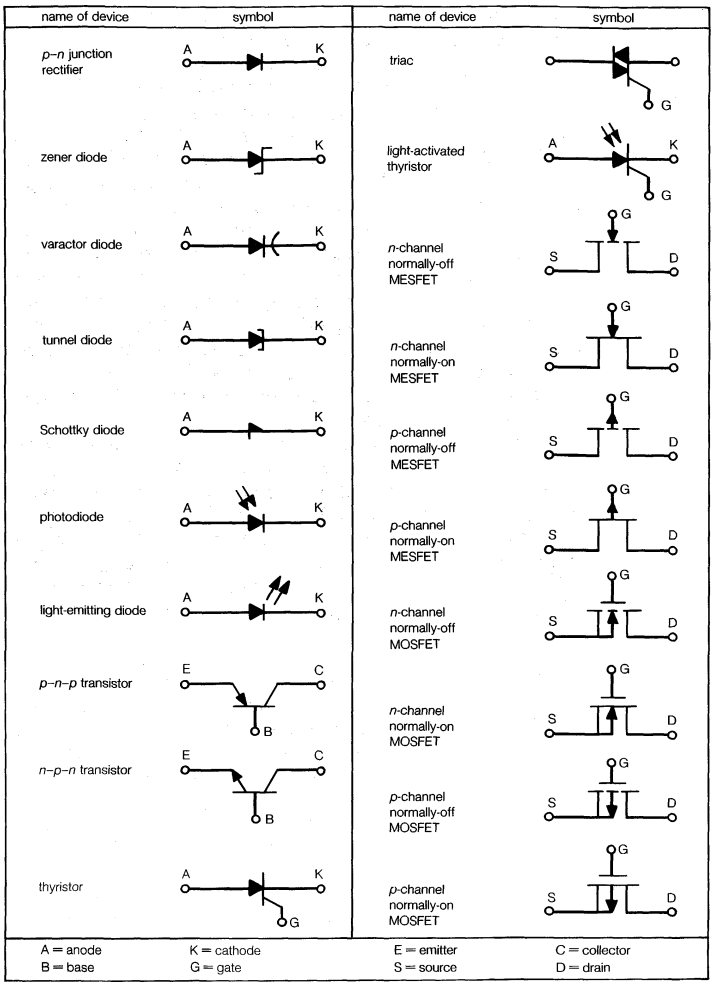


Figure 17: Names and circuit symbols for various semiconductor devices.

Zener diode. This voltage regulator is a *p–n* junction diode that has a precisely tailored impurity distribution to provide a well-defined breakdown voltage. It can be designed to have a breakdown voltage over a wide range from 0.1 volt to thousands of volts. The Zener diode is operated in the reverse direction to serve as a constant voltage source, as a reference voltage for a regulated power supply, and as a protective device against voltage and current transients.

Varactor diode. The varactor (variable reactor) is a device whose reactance can be varied in a controlled manner with a bias voltage. It is a *p–n* junction with a special impurity profile, and its capacitance variation is very sensitive to reverse-biased voltage. Varactors are widely used in parametric amplification, harmonic generation, mixing, detection, and voltage-variable tuning applications.

High sensitivity to reverse-biased voltage

Tunnel diode. A tunnel diode consists of a single *p–n* junction in which both the *p* and *n* sides are heavily doped with impurities. The depletion layer is very narrow (about 100 angstroms). Under forward biases the electrons can tunnel or pass directly through the junction, producing a negative resistance effect (*i.e.*, the current decreases with increasing voltage). Because of its short tunneling time across the junction and its inherent low noise (random fluctuations either of current passing through a device or of voltage developed across it), the tunnel diode is used in

special low-power microwave applications, such as a local oscillator and a frequency-locking circuit.

Schottky diode. Such a diode is one that has a metal–semiconductor contact (*e.g.*, an aluminum layer in intimate contact with a *n*-type silicon substrate). It is named for the German physicist Walter H. Schottky, who in 1938 explained the rectifying behaviour of this kind of contact. The Schottky diode is electrically similar to a *p–n* junction, though the current flow in the diode is due primarily to majority carriers having an inherently fast response. It is used extensively for high-frequency, low-noise mixer and switching circuits. Metal–semiconductor contacts can also be non-rectifying; *i.e.*, the contact has a negligible resistance regardless of the polarity of the applied voltage. Such a contact is called an ohmic contact.

Ohmic contacts

All semiconductor devices as well as integrated circuits need ohmic contacts to make connections to other devices in an electronic system.

The p–i–n diode. A *p–i–n* diode is a *p–n* junction with an impurity profile tailored so that an intrinsic layer, the "*i* region," is sandwiched between a *p* layer and an *n* layer. The *p–i–n* diode has found wide application in microwave circuits. It can be used as a microwave switch with essentially constant depletion-layer capacitance (equal to that of a parallel-plate capacitor having a distance between the plates equal to the *i*-region thickness) and high power-handling capability.

Transferred-electron diode. The transferred-electron diode (TED) is made from *n*-type compound semiconductors (such as gallium arsenide and indium phosphide [InP]) with two ohmic contacts. When an electric field is applied across the diode, it shows a negative resistance due to a field-induced transfer of electrons from high-mobility, lower energy levels to low-mobility, higher energy levels in the conduction band. TEDs have been used as local oscillators and power amplifiers, which cover the microwave frequency range from 1 to 100 gigahertz. They have become important solid-state microwave sources for use in radar, intrusion alarms, and microwave test instruments.

IMPATT diode. The IMPATT (impact ionization avalanche transit time) diode is another important solid-state source of microwave power. It employs impact ionization (a phenomenon related to reverse junction breakdown) and transit-time properties (*i.e.*, charge carriers travel across the depletion layer of a *p–n* junction) to produce negative resistance at microwave frequencies. The negative resistance of the type found in IMPATT diodes can be obtained from a conventional *p–n* junction, a Schottky diode, and many junction diodes that have special impurity profiles. Among solid-state devices, the IMPATT diode can generate the highest continuous-wave power output at millimetre-wave frequencies (*i.e.*, above 30 gigahertz).

Use of impact ionization and transit-time properties

Bipolar transistors. This type of transistor is one of the most important of the semiconductor devices. It is a bipolar device in that both electrons and holes are involved in the conduction process. The bipolar transistor delivers a change in output voltage in response to a change in input current. The ratio of these two changes has resistance dimensions and is a "transfer" property (input-to-output), hence the name transistor.

A perspective view of a silicon *p–n–p* bipolar transistor is shown in Figure 18A. Basically the bipolar transistor is fabricated by first forming an *n*-type region in the *p*-type substrate; subsequently a p^+ region (very heavily doped *p*-type) is formed in the *n* region. Ohmic contacts are made to the top p^+ and *n* regions through the windows opened in the oxide layer (an insulator) and to the *p* region at the bottom.

An idealized, one-dimensional structure of the bipolar transistor, shown in Figure 18B, can be considered as a section of the device along the dashed lines in Figure 18A. The heavily doped p^+ region is called the emitter, the narrow central *n* region is the base, and the *p* region is the collector. The circuit arrangement illustrated in Figure 18B is known as a common-base configuration. The arrows indicate the directions of current flow under normal operating conditions—namely, the emitter–base junction is forward biased and the base–collector junction is reverse

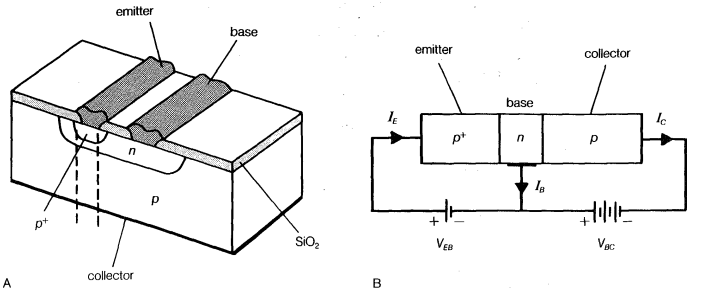


Figure 18: (A) Perspective of a *p–n–p* bipolar transistor; (B) idealized one-dimensional transistor.

biased. The complementary structure of the *p–n–p* bipolar transistor is the *n–p–n* bipolar transistor, which is obtained by interchanging *p* for *n* and *n* for *p* in Figure 18A. The current flow and voltage polarity are all reversed. The circuit symbols for *p–n–p* and *n–p–n* transistors are given in Figure 17.

The bipolar transistor is composed of two closely coupled *p–n* junctions. The emitter–base p^+–*n* junction is forward biased and has low resistance. The majority carriers (holes) in the p^+-emitter are injected (or emitted) into the base region. The base–collector *n–p* junction is reverse biased. It has high resistance, and only a small leakage current will flow across the junction. If the base width is sufficiently narrow, however, most of the holes injected from the emitter can flow through the base and reach the collector. This transport mechanism gives rise to the prevailing nomenclature: emitter, which emits or injects carriers, and collector, which collects these carriers injected from a nearby junction.

The current gain for the common-base configuration is defined as the change in collector current divided by the change in emitter current when the base-to-collector voltage is constant. Typical common-base current gain in a well-designed bipolar transistor is very close to unity.

The most useful amplifier circuit is the common-emitter configuration, as shown in Figure 19A, in which a small change in the input current to the base requires little power but can result in much greater current in the output circuit. A typical output current-voltage characteristic for the common-emitter configuration is shown in Figure 19B, where the collector current I_C is plotted against the emitter–collector voltage V_{EC} for various base currents. A numerical example is provided using Figure 19B. If V_{EC} is fixed at five volts and the base current I_B is varied from 10 to 15 microamperes (μA; 1 μA = 10^{-6} A), the collector current I_C will change from about four to six milliamperes (mA; 1 mA = 10^{-3} A), as can be read from the left axis. Therefore an increment of 5 μA in the input-base current gives rise to an increment of 2 mA in the output circuit—an increase of 400 times, with the input signal thus being substantially amplified. In addition to their use as amplifiers, bipolar transistors are key components for oscillators and pulse and switching circuits, as well as for high-speed integrated circuits. There are more than 45,000 types of bipolar transistors for low-frequency operation,

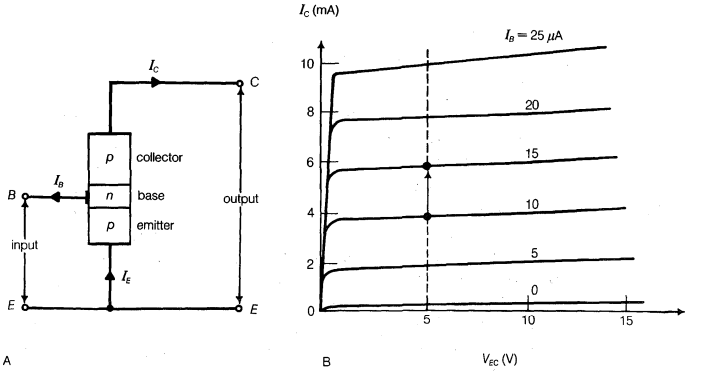


Figure 19: (A) Common-emitter configuration of a *p–n–p* transistor; (B) output characteristics for a *p–n–p* transistor in the common-emitter configuration.

with power outputs up to 3,000 watts and a current rating of more than 1,000 amperes. At microwave frequencies, bipolar transistors have power outputs of more than 200 watts at one gigahertz and about 10 watts at 10 gigahertz.

Thyristors. The thyristors constitute a family of semiconductor devices that exhibit bistable characteristics and can be switched between a high-resistance, low-current "off" state and a low-resistance, high-current "on" state.

Similarity to the bipolar transistor

The operation of thyristors is intimately related to the bipolar transistor, in which both electrons and holes are involved in the conduction processes. The name thyristor is derived from the electron tube called the gas thyratron, since the electrical characteristics of both devices are similar in many respects. Because of their two stable states (on and off) and low power dissipations in these states, thyristors are used in applications ranging from speed control in home appliances to switching and power conversion in high-voltage transmission lines. More than 40,000 types of thyristors are available, with current ratings from a few milliamperes to more than 5,000 amperes and voltage ratings extending to 900,000 volts.

Figure 20A provides a perspective view of a thyristor structure. An *n*-type wafer is generally chosen as the starting material. Then, a diffusion step is used to form the *p*1 and *p*2 layers simultaneously by diffusing the wafer from both sides. (Diffusion is the movement of impurity atoms into the crystalline structure of a semiconductor.) Finally, *n*-type impurity atoms are diffused through a ring-shaped window in an oxide into the *p*2 region to form the *n*2 layer.

A cross section of the thyristor along the dashed lines is shown in Figure 20B. The thyristor is a four-layer *p–n–p–n* diode with three *p–n* junctions in series. The contact electrode to the outer *p* layer (*p*1) is called the anode and that to the outer *n* layer (*n*2) is designated the cathode. An additional electrode, known as the gate electrode, is connected to the inner *p* layer (*p*2).

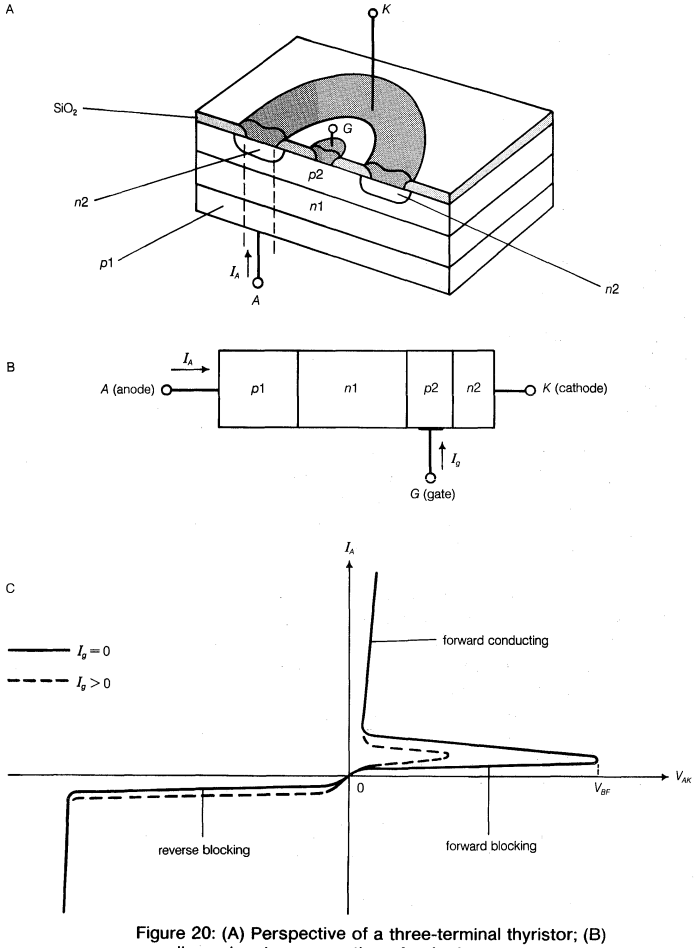


Figure 20: (A) Perspective of a three-terminal thyristor; (B) one-dimensional cross section of a thyristor; (C) current-voltage characteristics of a thyristor.

The basic current-voltage characteristic of a thyristor is illustrated in Figure 20C. It exhibits three distinct regions: the forward-blocking (or off) state, the forward-conducting (or on) state, and the reverse-blocking state, which is similar to that of a reverse-biased *p–n* junction. Thus, a thyristor operated in the forward region is a bistable device that can switch from a high-resistance, low-current off state to a low-resistance, high-current on state, or vice versa.

In the forward off state most of the voltage drops across the centre *n*1–*p*2 junction, while in the forward on state all three junctions are forward biased. The forward current-voltage characteristic can be explained using the method of a two-transistor analog, that is, to consider the device as a *p–n–p* transistor and an *n–p–n* transistor connected with the base of one transistor (*n*1) attached to the collector of the other. As the voltage V_{AK} in Figure 20C increases from zero, the current I_A will increase. This in turn causes the current gains of both transistors to increase. Because of the regenerative nature of these processes, switching eventually occurs and the device is in its on state. The maximum forward voltage that can be applied to the device prior to switching is called the forward-breakover voltage V_{BF}. The magnitude of V_{BF} depends on the gate current. Higher gate currents cause the current I_A to increase faster, enhance the regeneration process, and switch at lower breakover voltages. The effect of gate current on the switching behaviour is shown in Figure 20C (dotted line).

Triac

A bidirectional, three-terminal thyristor is called a triac. This device can switch the current in either direction by applying a small current of either polarity between the gate and one of the two main terminals. The triac is fabricated by integrating two thyristors in an inverse parallel connection. It is used extensively in AC applications such as light dimming, motor-speed control, and temperature control. There also are many light-activated thyristors that use an optical signal to control the switching behaviour of devices. The circuit symbols for some devices in the thyristor family are given in Figure 17.

Metal-semiconductor field-effect transistors. The metal-semiconductor field-effect transistor (MESFET) is a unipolar device because its conduction process involves predominantly only one kind of carrier. The MESFET offers many attractive features for applications in both analog and digital circuits. It is particularly useful for microwave amplifications and high-speed integrated circuits, since it can be made from semiconductors with high electron mobilities (*e.g.*, gallium arsenide, whose mobility is five times that of silicon). Because the MESFET is a unipolar device, it does not suffer from minority-carrier effects and so has higher switching speeds and higher operating frequencies than do bipolar transistors.

A perspective view of a MESFET is given in Figure 21A. It consists of a conductive channel with two ohmic contacts, one acting as the source and the other as the drain. The conductive channel is formed in a thin *n*-type layer supported by a high-resistivity semi-insulating (non-conducting) substrate. When a positive voltage is applied to the drain with respect to the source, electrons flow from the source to the drain. Hence, the source serves as the origin of the carriers and the drain serves as the sink. The third electrode, the gate, forms a rectifying metal–semiconductor contact with the channel. The shaded area underneath the gate electrode is the depletion region of the metal–semiconductor contact. An increase or decrease of the gate voltage with respect to the source causes the depletion region to expand or shrink; this in turn changes the cross-sectional area available for current flow from source to drain. The MESFET thus can be considered a voltage-controlled resistor.

A typical current-voltage characteristic of a MESFET is shown in Figure 21B, where the drain current I_D is plotted against the drain voltage V_D for various gate voltages. For a given gate voltage (*e.g.*, $V_G = 0$) the drain current initially increases linearly with drain voltage, indicating that the conductive channel acts as a constant resistor. As the drain voltage increases, however, the cross-sectional area of the conductive channel is reduced, causing an increase in the channel resistance. As a result, the current increases at a slower rate and eventually saturates. At a given drain

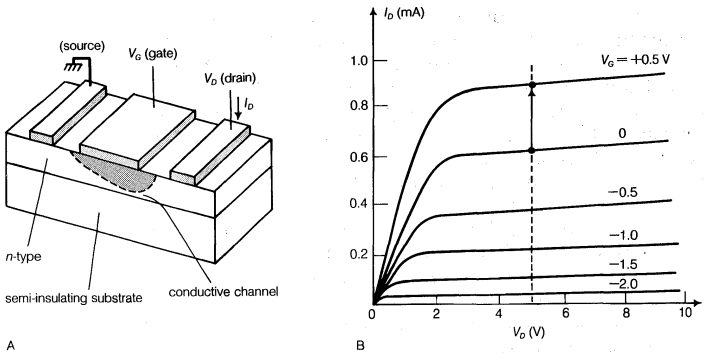


Figure 21: (A) Perspective of a MESFET; (B) current-voltage characteristic of a MESFET.

voltage the current can be varied by varying the gate voltage. For example, for $V_D = 5$ V, one can increase the current from 0.6 to 0.9 mA by forward biasing the gate to 0.5 V, as shown in Figure 21B; or one can reduce the current from 0.6 to 0.2 mA by reverse biasing the gate to −1.0 V.

A device related to the MESFET is the junction field-effect transistor (JFET). The JFET, however, has a *p–n* junction instead of a metal–semiconductor contact for the gate electrode. The operation of a JFET is identical to that of a MESFET.

There are basically four different types of MESFET (or JFET), depending on the type of the conductive channel. If, at zero gate bias, a conductive *n* channel exists and a negative voltage has to be applied to the gate to reduce the channel conductance, as shown in Figure 21B, then the device is an *n*-channel "normally-on" MESFET. If the channel conductance is very low at zero gate bias and a positive voltage must be applied to the gate to form an *n* channel, then the device is an *n*-channel "normally-off" MESFET. Similarly, *p*-channel normally-on and *p*-channel normally-off MESFETs are available. For the circuit symbols of these devices, see Figure 17.

To improve the performance of the MESFET, various heterojunction field-effect transistors (FETs) have been developed. A heterojunction is a junction formed between two dissimilar semiconductors, such as the binary compound GaAs and the ternary compound $Al_xGa_{1-x}As$.

Hetero-junctions

Such junctions have many unique features that are not readily available in the conventional *p–n* junctions discussed previously.

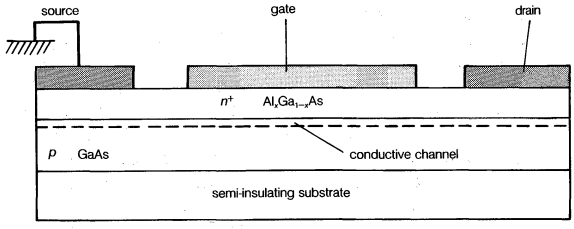


Figure 22: Cross section of a heterojunction FET having a conductive channel at the heterojunction interface.

Figure 22 shows a cross section of a heterojunction FET. The heterojunction is formed between a high-bandgap semiconductor (*e.g.*, $Al_{0.4}Ga_{0.6}As$, with a bandgap of 1.9 eV) and one of a lower bandgap (*e.g.*, GaAs, with a bandgap of 1.42 eV). By proper control of the bandgaps and the impurity concentrations of these two materials, a conductive channel can be formed at the interface of the two semiconductors, as indicated in Figure 22. Because of the high conductivity in the conductive channel, a large current can flow through it from source to drain. When a gate voltage is applied, the conductivity of the channel will be changed by the gate bias, which results in a change of drain current. The current-voltage characteristics are similar to those of the MESFET shown in Figure 21B. If the lower-bandgap semiconductor is a high-purity material, the mobility in the conductive channel will be high. This in turn can give rise to higher operating speed.

Metal-oxide-semiconductor field-effect transistors. The metal-oxide-semiconductor field-effect transistor (MOSFET) is the most important device for very-large-scale integrated circuits (those that contain more than 100,000 semiconductor devices such as diodes and transistors). The MOSFET is a member of the family of field-effect transistors, which includes the aforementioned MESFET and JFET.

A perspective view for an *n*-channel MOSFET is shown in Figure 23. Although it looks similar to a MESFET, there are four major differences: (1) the source and drain of a MOSFET are rectifying *p–n* junctions instead of ohmic contacts; (2) the gate is a metal-oxide-semiconductor structure, meaning that there is an insulator, silicon dioxide (SiO_2), sandwiched between the metal electrode and the semiconductor substrate, while for the MESFET the gate electrode forms a metal–semiconductor contact; (3) the left edge of the gate electrode must be aligned or overlapped with the source contact to facilitate device operation, while in a MESFET there is no overlapping of gate and source contact; and (4) the MOSFET is a four-terminal device, so that there is a fourth substrate contact in addition to the source, drain, and gate electrode, as in the case of a MESFET.

Differences between MOSFETs and MESFETs

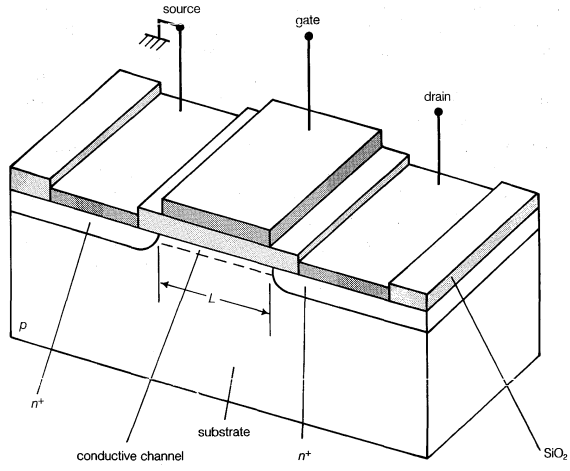


Figure 23: Perspective of a MOSFET.

One of the key device parameters is the channel length, *L*, which is the distance between the two n^+–*p* junctions, as indicated in Figure 23. When the MOSFET was first developed, in 1960, the channel length was longer than 20 micrometres (μm). Today, channel lengths less than I μm have been fabricated in volume productions, and lengths less than 0.1 μm have been created in research laboratories.

The source is generally used as the voltage reference and is grounded. When no voltage is applied to the gate, the source-to-drain electrodes correspond to two *p–n* junctions connected back to back. The only current that can flow from source to drain is a small leakage current. When a high positive bias is applied to the gate, a large number of electrons will be attracted to the semiconductor surface and form a conductive layer just underneath the oxide. The n^+ source and n^+ drain are now connected by a conducting surface *n* layer (or channel) through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltages; the conductance also can be changed by the substrate bias.

The current-voltage characteristic of a MOSFET is similar to that shown in Figure 21B. There are also four different kinds of MOSFETs, depending on the type of conducting layer. The four are *n*-channel normally-off, *n*-channel normally-on, *p*-channel normally-off, and *p*-channel normally-on MOSFETs. They are similar to MESFET varieties. The circuit symbols of these devices are given in Figure 17.

Types of MOSFETs

The main reasons why the MOSFET has surpassed the bipolar transistor and become the dominant device for very-large-scale integrated circuits are (1) the MOSFET can be easily scaled down to smaller dimensions, (2)

it consumes much less power, and (3) it has relatively simple processing steps, which results in a high manufacturing yield (*i.e.*, the ratio of good devices to the total). MOSFETs also are becoming important devices for power applications. This is because the MOSFET has a negative temperature coefficient at high current levels—*i.e.*, the current decreases as temperature increases. This characteristic leads to a more uniform temperature distribution over the device area and reduces MOSFET failures due to the localized heating that occurs in the bipolar transistor.

(S.M.Sz.)

INTEGRATED CIRCUITS

An integrated circuit is an assembly of electrically isolated circuit elements, both active semiconductor devices (transistors and diodes) and passive components (capacitors and resistors), together with electrically conducting interconnections, that are fabricated in place by an iterative process of lithographic definition, deposition, and etching on a common substrate (in most cases, silicon) in such a manner that the resulting interconnected elements perform an electrical circuit function. Many ICs, typically about 0.5 square centimetre each, are fabricated simultaneously on silicon wafers up to 20 centimetres in diameter and subsequently sawed into individual chips (dies) prior to packaging. An IC thus produced is usually sealed in a plastic package with electrical leads that are internally connected by fine wires to output pads on the silicon die and that permit the packaged IC to be plugged into a circuit card.

The integration of a large number of semiconductor devices on a single die of silicon is made possible by the high operational efficiency of the individual devices. Because of this, power dissipation is minimal and so too are the requirements for heat removal. The result is an IC of high dependability, which is further enhanced by the process of integration itself because the method of manufacturing the electrical interconnections between the devices and circuit elements by metal deposition and etching yields an extremely reliable circuit structure. The number of devices integrated on a single tiny chip has increased from an initial few to nearly 1,000,000 as circuit elements with ever-smaller features are employed. This has, in turn, led to a progressive increase in complexity, as exemplified by the dynamic random-access memory (DRAM) and logic circuits based on metal-oxide-semiconductor (MOS) technology (see Figure 24).

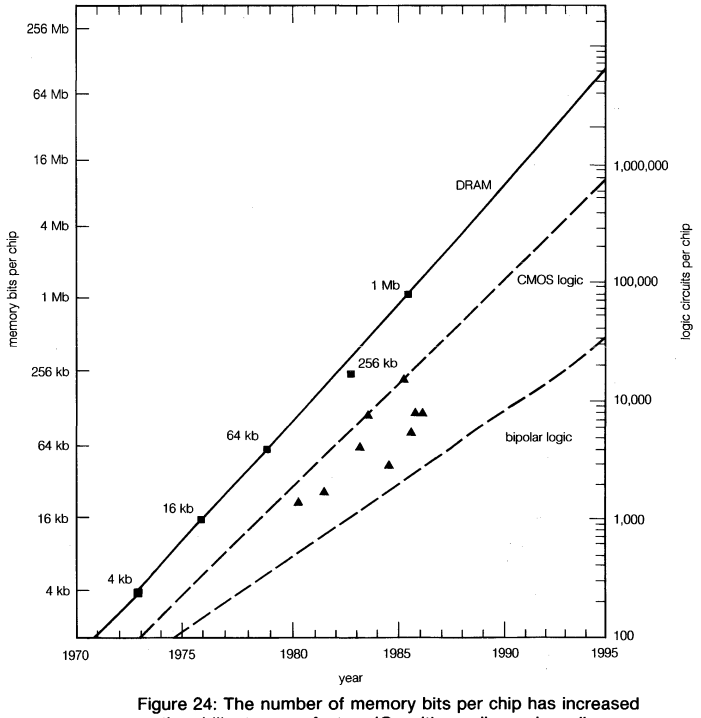


Figure 24: The number of memory bits per chip has increased as the ability to manufacture ICs with smaller and smaller feature sizes has improved. Likewise, the number of logic circuits per chip has increased. Because of the random nature of logic circuits, their density is not as great. Bipolar circuit density is less because of the larger bipolar circuit element and the power dissipation limitation on the number of circuits per chip.

The increase in complexity and resulting functionality has led to a dramatic decrease in cost per transistor and an increase in circuit performance. There are two reasons for the latter: (1) the performance of the individual active devices improves the smaller their internal dimensions become; and (2) the quality of the performance of the entire circuit improves the closer the active devices are positioned to one another. Large-volume applications for ICs have further lowered the manufacturing costs per IC. The cumulative result is that the integrated circuit has become the most pervasive technology of the 20th century. It has provided the cornerstone of modern microelectronics and has promoted the development of the so-called information society. Applications of ICs range from their use in supercomputers, which are bringing about revolutionary advances in medical diagnosis, biotechnology, aeronautical and space engineering, telecommunications, and defense systems, to the development of new consumer products capable of bringing services and information to the home and office environment that otherwise would not have been possible.

Diverse applications of ICs

Device/circuit technology. There are two basic types of transistors that are extensively employed in integrated circuits. They are the bipolar transistor and the MOSFET. (For a detailed discussion of these devices, see above *Bipolar transistors* and *Metal-oxide-semiconductor field-effect transistors.*)

The bipolar transistor is a current-controlled device. As voltage is applied to the base, injecting current into the region, the collector current increases. The ratio of collector current to injected base current is greater than unity, resulting in gain. The transistor can therefore be used as a high-gain amplifier in analog applications. It also should be noted that, for a fixed base current, the collector current saturates and is relatively insensitive to changes in emitter–collector voltage (see Figure 19). Thus, the transistor can also be used as a current switch in digital-circuit applications. The output impedance of the device is low, so that the transistor behaves as a current source and is thus able to drive capacitive loads.

The MOSFET is a voltage-controlled device. As voltage is applied to the gate, conduction begins between the source and the drain, with a consequent voltage gain (see Figure 23). At a fixed gate voltage the drain current saturates and is insensitive to changes in the drain voltage. The output impedance of the device is high, so that the transistor functions as a voltage source. It dissipates much less power than the bipolar transistor and therefore is more suitable for integration than the latter, especially for digital-circuit switching applications. The bipolar device, however, is more proficient in analog and current-driver applications. The MOSFET and bipolar transistor complement each other in the two principal types of integrated-circuit applications, analog and digital.

Analog circuits. These circuits are designed to respond to an analog input and produce an analog output where information is conveyed by the instantaneous value of the signal. The principal technology for analog applications has been bipolar, largely because bipolar technology generates less noise and allows greater small-signal gain. MOS digital technology, however, is becoming more widely used for analog applications as a result of the increasing capability of digital MOSFET integrated circuits to handle complex functions. Special circuit techniques have been developed to overcome some of the shortcomings of digital MOS devices for such applications. Charge storage on integrated capacitors is one such technique, made possible by the high-input impedance and zero offset of MOS technology. Arrays of integrated capacitors are used to sample and hold the analog signal, permitting it to be digitally encoded. On the other hand, in order to furnish an analog interface to sensors and other "real world" input/output devices for computers and computer-based equipment, bipolar analog functions are being integrated on MOS

Growing use of MOS technology

chips that are largely digital functions. Additionally, bipolar technology is being integrated because its greater capability to drive current is useful for line-driver and output-driver transistors. For very-large-scale integrated circuits a mixed bipolar/complementary MOSFET technology (BIC-MOS) benefits from the advantages of both bipolar and MOSFET technologies within the same integrated circuit. Typical analog applications for integrated circuits include their use as filters, analog-to-digital and digital-to-analog converters, voltage comparators and regulators, and operational amplifiers.

Operational amplifiers constitute the basic building block for analog circuits. The ideal operational amplifier has infinite input impedance, infinite gain, and zero output impedance. Although bipolar-circuit designers have not achieved this ideal, they have approached it closely enough for most practical applications by relying heavily on signal-feedback techniques. A very popular implementation of bipolar technology for general-purpose operational-amplifier application is the circuit shown in Figure 25. A high-impedance input is achieved by the emitter-follower input transistors, T1 and T2, whose output drives the emitters of T3 and T4, configured as a common-base differential amplifier pair. T5 and T6 together constitute an active load for the differential pair. Additional stages supply voltage gain and emitter-follower output to minimize loading and provide for a low-impedance output. A total of 20 bipolar transistors and 11 resistors are integrated onto a single silicon die. The operational amplifier is available for commercial use in an eight-pin plastic package and is characterized for operation over a temperature range of 0 to 70° C.

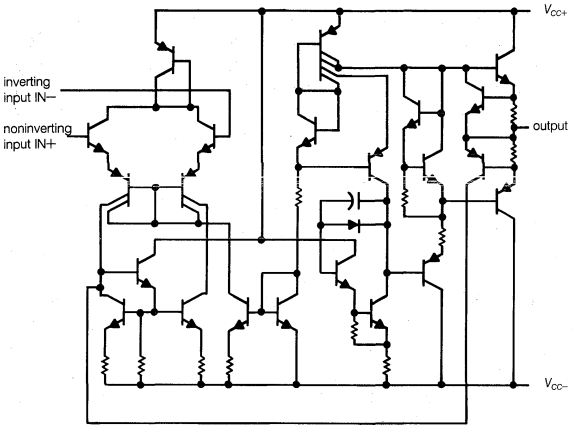


Figure 25: A general-purpose operational amplifier, the uA741C.

Digital circuits. These circuits, based principally on MOS technology, have played the major role in the growth of microelectronics since the concept of the IC was developed. The transistors of a digital circuit are operated as switches having only two states: on and off. They are combined in a manner to perform logic, arithmetic, and memory applications. All four types of MOSFETs, enhancement and depletion *p*-channel metal-oxide-semiconductor (PMOS) transistors and enhancement and depletion *n*-channel metal-oxide-semiconductor (NMOS) transistors, have been used in MOS digital-circuit design as the technology has evolved. Initially PMOS circuits were exclusively employed because of the greater ease of high-yield manufacturing. Later, as manufacturing procedures improved (specifically, airborne particulate and impurity contamination was minimized), NMOS circuits evolved because of their improved performance relative to PMOS circuits. More recently, a complementary MOS (CMOS) technology employing both PMOS and NMOS transistors superseded both previous circuit technologies because of its reduction in power dissipation.

CMOS technology

The algebra for logic design with binary quantities, "on" and "off" representing 1 and 0 (*i.e.*, true and false), was invented by the English mathematician and logician George Boole during the mid-19th century when he also published

A	$\bar{A}$
0	1
1	0

A	B	A · B
0	0	0
0	1	0
1	0	0
1	1	1

A	B	A+B
0	0	0
0	1	1
1	0	1
1	1	1

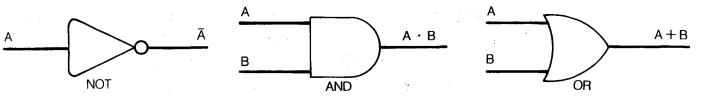


Figure 26: The logic symbol, its corresponding function, and the truth table defining the operation are shown. The NOT function inverts the signal (*i.e.*, a 1 becomes a 0 and a 0 becomes a 1). The AND function generates a true, or 1, if both inputs are 1; otherwise the output is false, or 0. The OR function generates a 1, or true, if either input is a 1, or true, value.

Adapted from W.C. Holton, "The Large-Scale Integration of Microelectronic Circuits"; copyright © 1977 by Scientific American, Inc.; all rights reserved

his ideas. The three basic logic functions "NOT," "AND," and "OR" of Boole's system, together with their truth tables, are given in Figure 26. (Arithmetic is performed in the binary number system employing Boolean logic.) The circuit implementation of these functions in both NMOS and CMOS technologies is shown in Figure 27. These basic elements are combined in the design of integrated circuits (for digital computers and associated devices) to perform the desired functions. As an illustration, the circuit for the binary addition of a single place of two binary numbers is shown in Figure 28.

In addition to logic and arithmetic circuits, memory

Adapted from Amr Mohsen in Norm Einspruch (ed.), *VLSI Handbook* (1985); Academic Press

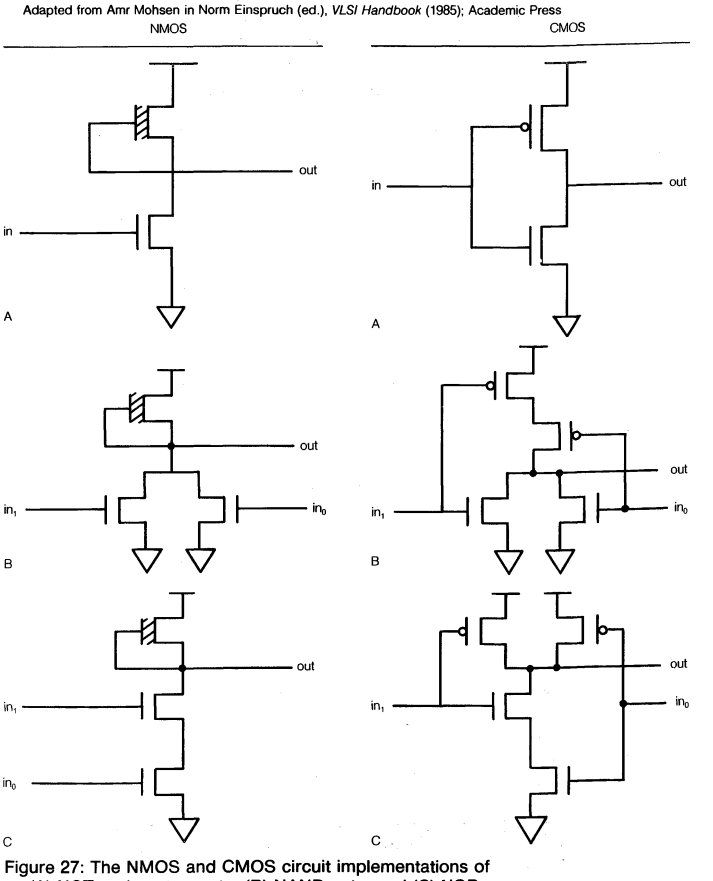


Figure 27: The NMOS and CMOS circuit implementations of an (A) NOT, or inverter, gate, (B) NAND gate, and (C) NOR gate. The NOT gate is attached to the outputs of the NAND and NOR gates, respectively, to invert the signal and achieve AND and OR gate logic functions. The small circle at the gate symbol for the MOSFET designates a *p*-channel device.

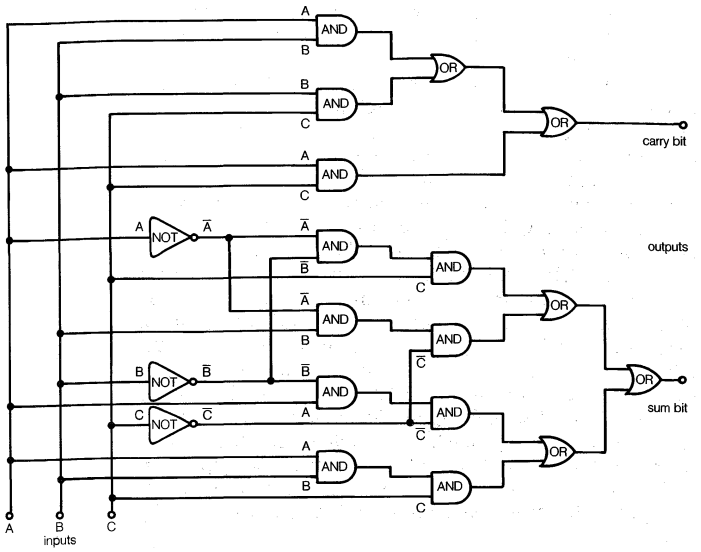


Figure 28: Array of logic gates that provides for the binary addition of a single place of two binary numbers, where the three inputs are a single place, A and B, of the two binary numbers and the carry bit, C, from the previous addition of the next lower place. The output is the sum bit and the carry bit to the next place in the addition of the two binary numbers.

Adapted from W.C. Holton, "The Large-Scale Integration of Microelectronic Circuits"; copyright © 1977 by Scientific American, Inc.; all rights reserved

DRAMs and other memory circuits

circuits that provide for the storage of information are primarily fabricated with MOS technology. Several types of memory circuits have been developed. The simplest and highest-density memories are the DRAMs. They are implemented by storing charge on a capacitor to represent 1 or 0. The charge is stored and sensed by a pass-gate MOSFET to write or read the memory capacitor cell (see Figure 29). The memory consists of an array of these cells, together with the logic circuits that permit random-access addressing of cells in the array. Because the capacitor slowly discharges (*i.e.,* loses its information), the memory must be "refreshed" frequently to restore the charge to its intended value, hence the name dynamic memory. A static memory, one that does not require refreshing, can be constructed from six transistors (see Figure 29). In this kind of memory the information is stored by the state of conduction—*i.e.,* either the right or left path of the cell is conducting. This state may be read or changed by addressing the cell. Both of these memory devices lose the stored information when the electrical power to the circuits is removed. Nonvolatile memories that retain their information by storing a charge on an electrically isolated conductor have been developed: the EPROM (electrically programmable read-only memory), in which the information is electrically written but erased only by the application of ultraviolet light on the IC, and the EEPROM (electrically erasable and programmable read-only memory), in which the information is both written and erased electrically. One other nonvolatile memory, the ROM (read-only memory), in which the information is written at the time of manufacture, provides permanent memory capability.

Adapted from D.A. Hodges and H.G. Jackson, *Analysis and Design for Digital Integrated Circuits* (1983); McGraw-Hill Book Co., New York

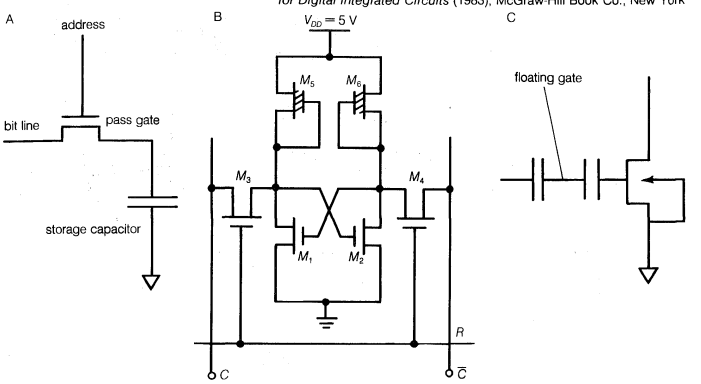


Figure 29: NMOS circuit implementation of (A) a DRAM cell, (B) an SRAM cell, and (C) an EPROM transistor, where the charge is stored on the floating gate. The address line selects the cell to be written or read, and the memory-state information is sensed on the bit line(s).

Very-large-scale integrated circuits perform higher-level functions by combining both logic and memory circuitry on a single silicon chip. Such devices include microprocessors, microcomputers, digital-signal processors, and application-specific integrated circuits for more customized uses. Representative of an IC of this complexity level is the digital signal processor shown in Figure 30.

Designing integrated circuits. Devices produced during the early 1960s, which marked the initial era of integrated circuits, contained fewer than 30 transistors and required only about 1,000 "geometries" (or geometric patterns; see below) to describe the location of the circuit elements on a silicon die. These relatively simple ICs were products of the so-called small-scale integration (SSI) era. By the mid-1980s, the era of very-large-scale integration (VLSI), the number of transistors per integrated circuit exceeded 1,000,000. Each of these ICs required more than 10,000,000 geometries to describe the design. The ever-increasing complexity of integrated circuits is such that computer-aided design (CAD) tools are now required to produce error-free, cost-effective designs.

Reliance on CAD systems

Texas Instruments Inc.

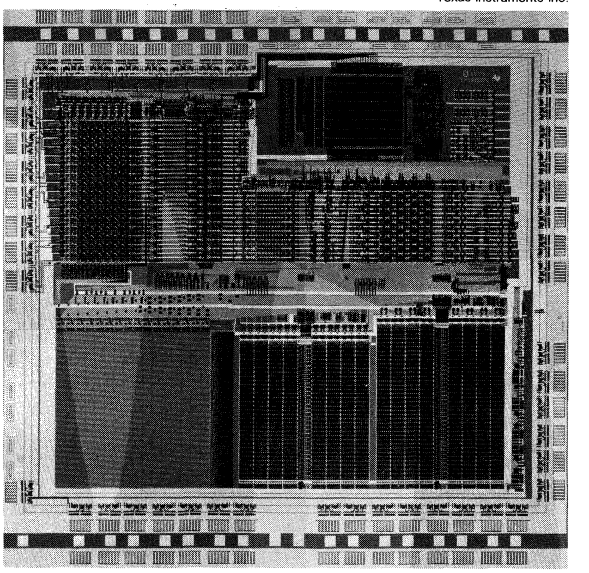


Figure 30: Texas Instruments Incorporated TMS320C25 digital signal processor.

The design process begins with the definition of an IC product or the function to be performed by it. From this definition, a system architect specifies the functional behaviour of the system—say, a supercomputer—and partitions the system into the functional elements needed. Some of these elements may be available from previous designs and are already described at all succeeding levels of design, not requiring further definition for the new design. Other elements, however, do need further defining of their internal operation at the "register-transfer level" to further define their functions. The specific logic description is extracted from this definition, and the design of the electrical circuit follows. This leads to the generation of a data base for the IC design that contains explicit information about the behaviour of the system in response to external stimuli; the architecture of the system, the logic elements, and electrical circuits; a timing analysis of the system's circuits; and all voltages and currents.

From this circuit-level description, designers prepare the physical layout of the geometries from which are produced the photomasks to be used in lithographic patterning during the IC manufacturing process (see below). Design rules specifying the minimum-allowed dimensions of the patterns, along with the device-level descriptions and the

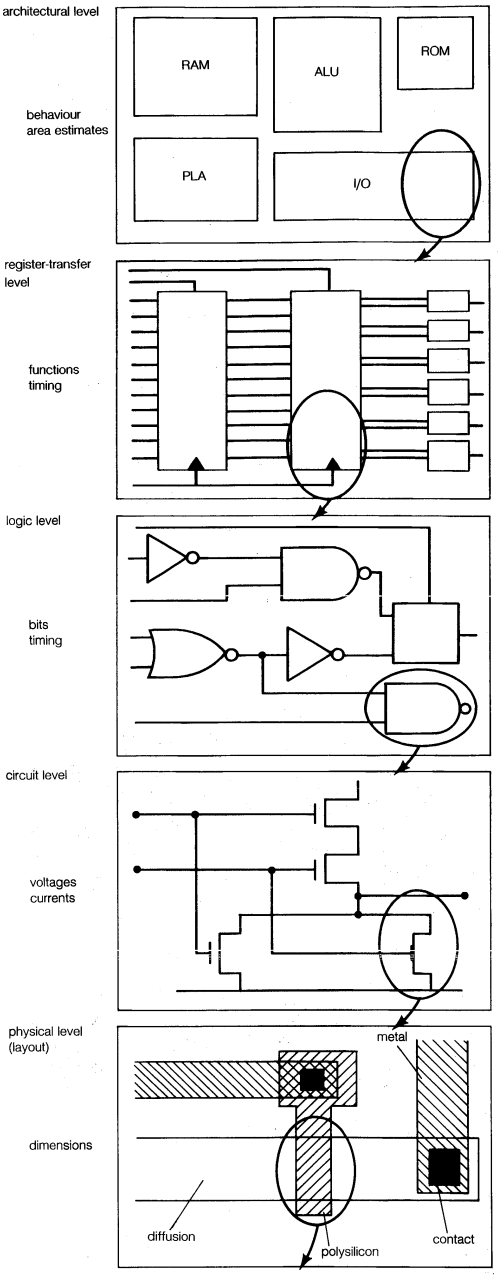


Figure 31: The IC design process.
Adapted from Amr Mohsen in Norm Einspruch (ed.), *VLSI Handbook* (1985); Academic Press

technology utilized in IC manufacture, are taken into account in producing the layout. The initial specification and the logic implementation provide the basis for test patterns that will be used to determine the functionality of the integrated circuit at the completion of its manufacture. The entire IC design process is depicted in Figure 31.

Manufacturing technology. The fabrication of integrated circuits begins with the preparation of silicon of very high purity. Single-crystal boules with a diameter as large as 20 centimetres are produced from the silicon. The boules are sliced into wafers of a specified crystal orientation. When their surfaces have been polished to a mirrorlike finish that is free of defects, the silicon wafers are ready for fabrication into IC devices.

Basic steps in IC fabrication

This manufacturing process typically involves a sequence of more than 200 specific steps. These steps may be categorized according to function: deposition of thin film, introduction of impurities (doping), lithographic patterning of IC features corresponding to those of the physical layout, etching to define the features of individual circuit elements, and cleaning to prevent contamination and the consequent introduction of defects into circuit elements by particulate matter. These basic operations are repeated again and again in different sequential order until the IC unit is completed.

Film deposition. Several kinds of thin films are deposited on a silicon wafer by different methods during various stages of the fabrication process. The initial step following cleaning is the formation of a silicon dioxide film. This film is grown by placing the silicon wafer in an oxidizing environment at high temperature, as, for example, in a quartz-walled furnace tube. This operation may be followed by the deposition of a film of silicon nitride. Later in the fabrication process, metal films are deposited by means of sputtering. In this technique, a cathode made of the material to be deposited is bombarded with positive ions, resulting in the ejection of atoms from the cathode. These atoms are deposited on the wafer surface and form the desired metal film. Metal and polysilicon films are formed by chemical vapour deposition. Various metal silicide films are deposited and reacted with the surface of the silicon wafer. Glass films also may be deposited, and photoresist films (those of a photosensitive material) are applied by dropping a liquid polymer onto a rapidly rotating wafer.

Impurity doping. Selected impurities (*e.g.*, boron and phosphorus) are introduced into the silicon substrate to control its conductivity in a selective manner. This is accomplished by two methods: ion implantation and thermal diffusion. In the ion implantation process the silicon wafer is exposed to a beam of energetic particles (*i.e.*, high-energy ions) of the substances that are to be incorporated into the silicon. These impurities are driven into the silicon wafer to a depth ranging from a few hundred angstroms to several micrometres, depending on the energy and the mass of the ions in the beam. The wafer is annealed after implantation of the impurity ions in order to eliminate any damage that may have been done to the silicon during the process.

In diffusional doping, specific regions of the silicon wafer are exposed to concentrations of the selected impurities under intense heat in a special high-temperature furnace. At temperatures of about 800° C or higher, the impurity atoms are able to enter the silicon crystal lattice and diffuse into the upper layers of the crystal.

Lithographic patterning. The process of lithographic patterning determines the geometric features specified by the layout as the integrated circuit is fabricated layer by layer. A photomask containing pattern information is prepared for each layer. An image of the photomask is projected onto the surface of the silicon wafer. This is commonly done with an optical projection printer after

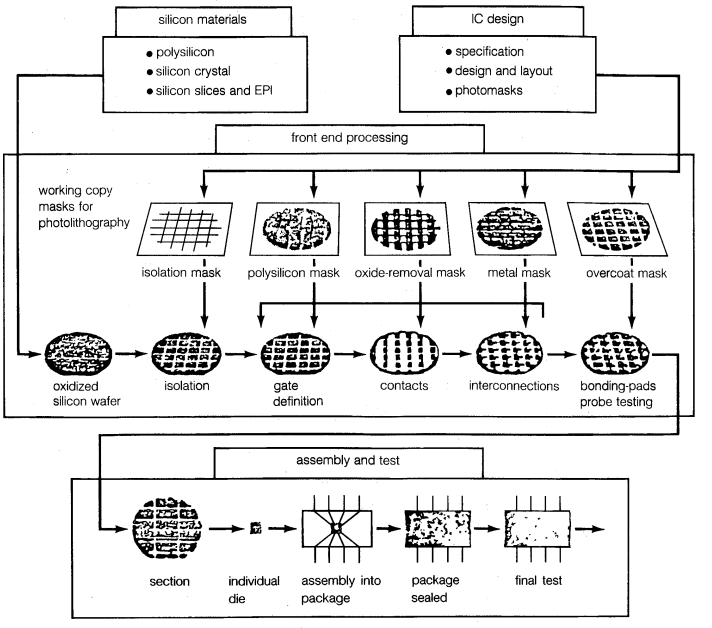


Figure 32: Typical IC design and manufacturing process for NMOS integrated circuits.

the wafer has been coated with a thin layer of photoresist. The pattern image is developed, and the exposed photosensitive material is removed chemically (in the case of a positive photoresist) from the areas that have been exposed to light. The desired pattern is thereby transferred from the photomask to the photoresist on the surface of the partially fabricated integrated circuit. This photoresist pattern serves to define the areas of the wafer where, during a subsequent process step, film is to be deposited, material is to be removed by etching, or impurities are to be introduced. Once the selective deposition or removal of material has been accomplished, the remaining photoresist is cleared off the wafer.

Etching. During this process, material is selectively removed from the wafer surface as defined by the patterned photoresist in order to define the structure of the previously deposited layer. The etching process is accomplished by exposing the wafer to a gas plasma, which both chemically reacts with the material to be removed and physically ablates it.

Manufacturing an NMOS circuit

A representative process sequence. The fabrication of an NMOS circuit can be used to illustrate how the process steps described above might be combined. The sequence of basic steps in manufacturing an NMOS circuit is shown in Figure 32.

Preliminary to these steps is the cleaning of the single-crystal silicon wafer. The fabrication process gets under way with oxidation and nitridation, which are followed by application of the photoresist and photolithographic definition of the mask pattern for electrical isolation between circuit components. After the photoresist has been developed and the isolation oxide/nitride mask defined by etching, the wafer is ion-implanted to electrically isolate the areas where the MOSFETs are to be formed, and oxide is grown over the implanted regions. The oxide/nitride film is then removed and a thin gate oxide is regrown in these regions. Polysilicon is deposited and patterned in the next photolithographic process using a second photomask to define the gate structure of the NMOS transistors. A subsequent ion-implantation process introduces the required impurities into the source/drain regions of the MOSFETs, and silicon oxide is deposited. The positions where electrical contact is to be made to the MOSFETs are defined, using the oxide-removal mask and an etch process. Metal is then deposited into the opened "vias" (passages) in the oxide layer and over its surface. During the subsequent photolithographic process, it is patterned to form the desired electrical interconnections. These two steps are repeated for each succeeding level to produce additional levels of interconnections. Finally, a protective overcoat of oxide/nitride is applied, and vias are opened so that the wires connecting the IC chip to its carrier package can be bonded to output pads.

Adapted from A.B. Glaser and G.E. Subak-Sharpe, *Integrated Circuit Engineering: Design, Fabrication, and Applications* (1977); Addison-Wesley Publishing, Reading, Mass.

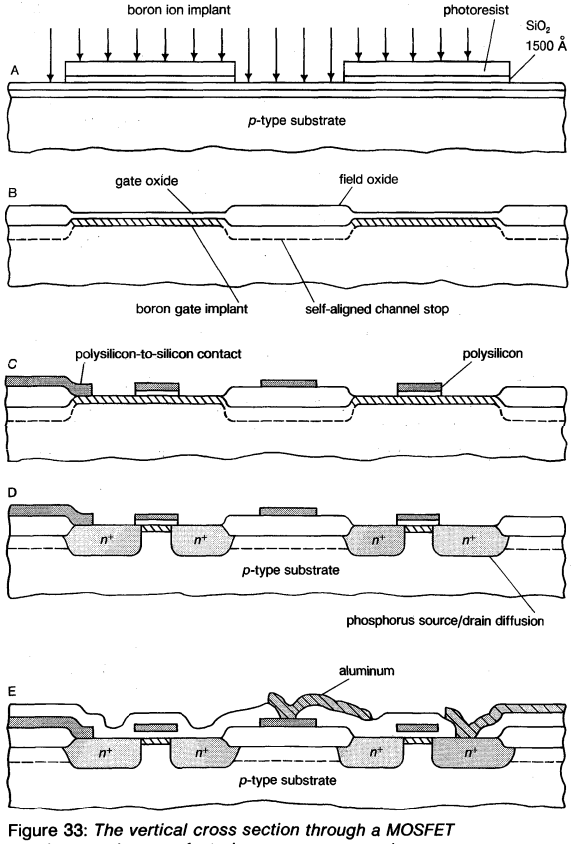


Figure 33: *The vertical cross section through a MOSFET transistor as the manufacturing process proceeds.*
(A) Following oxide/nitride growth, patterning, and isolation implantation; (B) after growth of isolation oxide and gate oxide for the MOSFETs; (C) following polysilicon deposition and patterning; (D) after source/drain formation; and (E) the completed structure with a single level of interconnect.

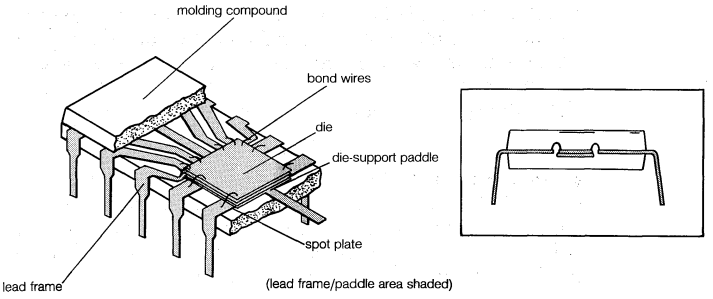


Figure 34: Integrated circuit (cutaway view) in molded plastic package.

The assembly process begins with an electrical test of the ICs die by die on each wafer to determine the location of the good units. The wafer is then sawed into individual dies, and each good unit is installed in a package (made either of plastic or ceramic) by attaching the die to a lead frame and soldering wires between the output pads on the die and the internal leads of the package. A final test is performed on the packaged die to determine whether the unit operates within the specified standards.

These steps, depicted schematically in Figure 32, illustrate the relationship between the design process, which determines the character of the photomasks, and the manufacturing process. Figure 33 shows the resulting vertical profile through a MOSFET structure as the transistor and subsequent electrical interconnections are manufactured. A cutaway view of a finished integrated circuit packaged in a molded plastic enclosure is given in Figure 34.

Possible developments. Silicon integrated-circuit technology, which is based on classical physics, is believed to be viable until the minimum feature size reaches below 0.1 micrometre where quantum mechanical effects begin to be important. However, device structures ranging from 0.01 to 0.1 micrometre, based on quantum domain phenomena, promise to supplant present-day IC technology. Also, other semiconductor materials, particularly those composed of III-V compounds (see above), have the physical properties and flexibility needed to produce novel devices capable of superior performance. Their suitability for integrated circuitry and the viability of their manufacturing techniques, however, remain to be proved. (W.C.Ho.)

OPTOELECTRONIC DEVICES

Major classes of optoelectronic devices

Optoelectronic devices are devices in which the photon, the basic particle of light, is affected. Such devices can be divided into four groups: (1) photodetectors and solar cells that convert photons into electrical current, (2) light-emitting diodes and semiconductor lasers that convert an applied voltage into emitted photons, (3) optical fibres that guide light within a small plastic or glass fibre between a light source and detector, and (4) liquid-crystal displays that use an applied voltage to change the reflection of light. A useful expression when dealing with optoelectronics is the conversion between the wavelength of the light, λ, in micrometres, and the energy of the photon, E, in electron volts:

$$\lambda = 1.24/E.$$

Photodetectors and solar cells. *Photodetectors.* A photodetector is a semiconductor device that transforms light into an electrical signal. Its applications range from sensing the amount of light for control of the shutter in an automatic camera to conversion of an optical signal in an optical communication system into an electrical signal. There are two basic types of photodetectors: photoconductors and photodiodes. Silicon is used for photodiodes in the 0.8- to 0.9-micrometre wavelength region, while germanium and III-V compound semiconductors are employed for those in the 1.0- to 1.6-micrometre wavelength region. At longer wavelengths, narrow-energy-gap compound semiconductors are used; they are cooled to the temperature of liquid nitrogen (77 K) to reduce the leakage current.

Photoconductors

A photoconductor is generally a thin semiconductor layer with ohmic contacts. The basic structure is shown in Figure 35. In this example, a high-resistivity substrate of gallium arsenide has highly doped n^+ regions for ohmic contact to the metal contacts and a lightly doped n^- region where photons with energy larger than the energy gap are absorbed.

The absorption of light is governed by the expression

$$I = I_o \exp(-\alpha x),$$

where I is the light intensity at the depth x, I_o is the incident light intensity, and α is the absorption coefficient at a particular photon energy. The absorption coefficients for commonly used semiconductors in the visible and near-infrared spectral region are shown in Figure 36.

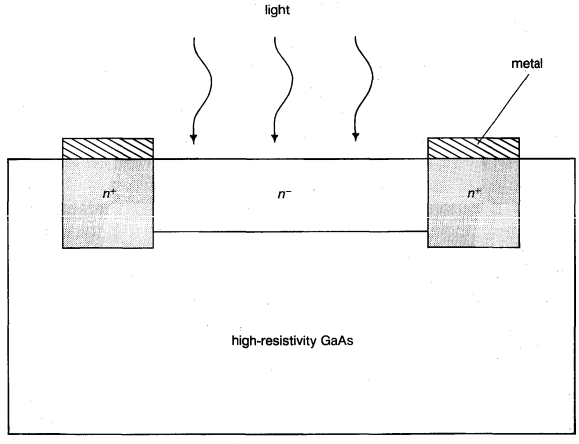


Figure 35: A photoconductor.

A small voltage is applied across the metal contacts, and the absorbed light varies the current that flows in the external circuit. The absorbed light generates free carriers to give the photocurrent I_p. The total number of absorbed photons, when multiplied by the charge q, gives the primary photocurrent I_{ph}. The photoconductor gain is given by

$$\text{gain} = I_p/I_{ph} = \tau/t_r,$$

where τ is the carrier lifetime and t_r is the carrier transit time. For small electrode spacing, the gain can significantly exceed unity. With high carrier mobility and short carrier lifetime, together with small electrode spacing, the sensitivity and speed of response can approach the results obtained with photodiodes. The advantage of photoconductors is the ease of fabrication.

Photodiodes

A photodiode is a p–n junction in which electron–hole pairs are created by photons absorbed near the p–n junction. A schematic representation is shown in Figure 37. These photo-generated carriers result in a current through the external circuit. The photodiode is generally intended to operate only over a narrow wavelength region of a particular optical signal source. The energy gap of the semiconductor should be slightly less than the photon energy. The photodiode can be operated in the photovoltaic mode in which no bias voltage is applied, but the device is connected to a load resistor. The optical signal then appears as a voltage across the load resistor. The diode

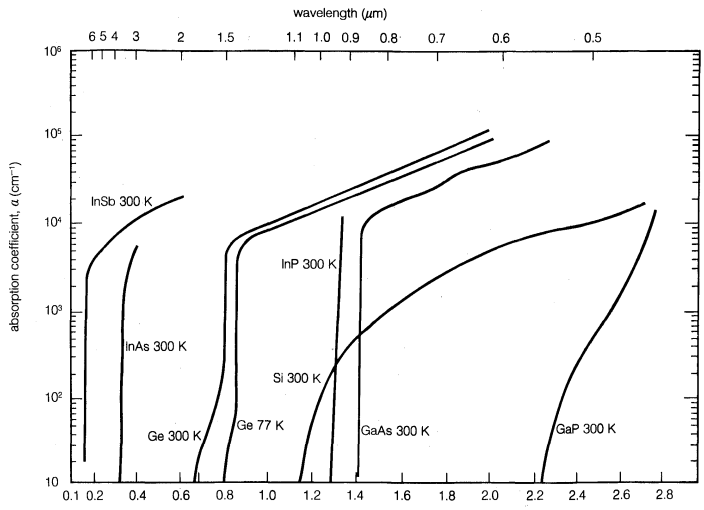


Figure 36: Absorption coefficients for silicon, germanium, and selected III-V compound semiconductors in the visible and near-infrared spectral region (see text).

From G.E. Stillman, V.M. Robbins, and N. Tabatabaie, "III-V Compound Semiconductor Devices: Optical Detectors," IEEE Transactions on Electron Devices, ED-31, 1643, © 1984 IEEE

area is small so that it minimizes the capacitance for better high-frequency response. An important measure of how well the device converts photons to electrons is the quantum efficiency, which is the ratio of the number of electrons flowing in the external circuit to the number of incident photons. With an antireflection coating, the quantum efficiency can exceed 90 percent.

One of the most commonly used photodiodes is the p–i–n diode, which has a very lightly doped i region between the p and n regions (see Figure 38). This structure can enhance the quantum efficiency because more of the light is absorbed in the diode depletion region (between the p and n regions), where it is collected and can flow through the load resistor. Also, because the p and n regions are farther apart than in the p–n-junction photodiode, the capacitance is less and the response speed is faster.

The avalanche photodiode is designed for use at high-reverse bias to provide current gain to the photogenerated hole–electron pairs. When the electric field in the depletion region of the reversed biased p–n-junction photodiode is about 10^5 volts per centimetre, a photogenerated hole or electron can collide with adjacent electron-bonding atoms, break the bond, and create a hole–electron pair. This process is called impact ionization. These newly created pairs can gain enough energy from the electric field to cause further impact ionization until finally an avalanche of carriers is produced. The avalanche gain can increase the signal-to-noise ratio and thus detect lower-intensity light signals than can other photodetectors. The avalanche photodiode, however, will have low noise if only the electron or hole is capable of causing impact ionization. Silicon has a large difference in the ionization rates of electrons

Impact ionization

From S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (1979); John Wiley and Sons, New York

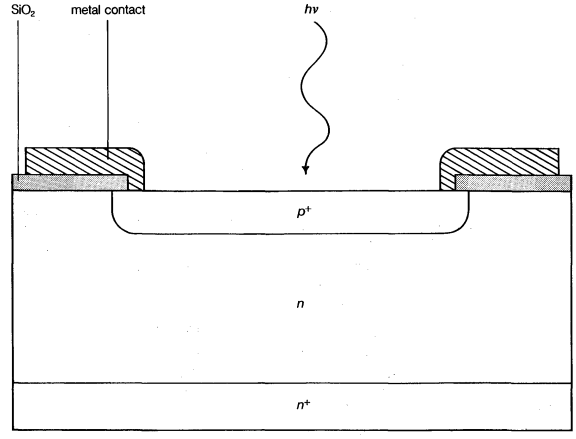


Figure 37: The p–n junction photodiode.

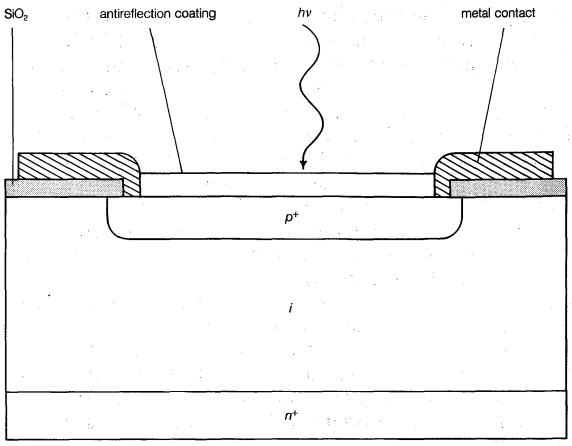


Figure 38: The *p–i–n* photodiode.
From S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (1979); John Wiley and Sons, New York

and holes, while the ionization rates for the III-V semiconductors are nearly equal. Thus only silicon has been extensively used for avalanche photodiodes.

Solar cells. The solar cell was the first optoelectronic device developed and was demonstrated by D.M. Chapin, C.S. Fuller, and G.L. Pearson in 1954 with a diffused silicon *p–n* junction. The solar cell is a large-area photodiode that "detects" the solar emission spectrum rather than a specific optical signal wavelength, as do photodiodes. The solar cell is unbiased, and the load is connected directly across the two terminals of the *p–n* junction. One of the most important parameters is the conversion efficiency, which is the ratio of the maximum power output to the incident power.

Power source for space satellites

The first space satellites were electrically powered by silicon solar cells, and these cells continue to be an important long-duration power source for satellites. The solar cells originally used for this purpose were made with single-crystal silicon and had conversion efficiencies of about 15 percent. For ground applications, lower-cost solar cells have been developed by using large-grained, polycrystalline silicon with efficiencies near 12 percent instead of the more expensive single-crystal silicon. Even thin-film amorphous solar cells with efficiencies of about 6 percent have been investigated for further cost reduction. In each case the lower cost is accompanied by reduced conversion efficiency. Other materials, such as aluminum gallium arsenide or gallium arsenide, have been used in applications where an increased conversion efficiency of more than 25 percent can justify the significantly greater cost.

The design of solar cells is influenced by the solar emission spectrum. The effect of the atmosphere on sunlight at the Earth's surface is defined by the air mass. The solar spectrum outside the atmosphere is the air mass zero (AM0). The solar spectrum at the Earth's surface for the minimum path length with the Sun directly overhead is AM1. The sunlight is attenuated by the atmosphere due to the absorption of infrared rays by ozone and scattering by clouds and airborne particles (*e.g.*, dust). For the Sun at an angle of 60° from the overhead position, the spectrum is AM2. The solar spectral irradiance is given by the power per unit wavelength. These three solar spectral irradiance spectra are shown in Figure 39, together with the 5,800 K emission spectrum of a blackbody that approximates the AM0 spectrum. (A blackbody is a hypothetical ideal body or surface that absorbs and reemits all radiant energy falling upon it.)

From S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (1979), John Wiley and Sons, New York

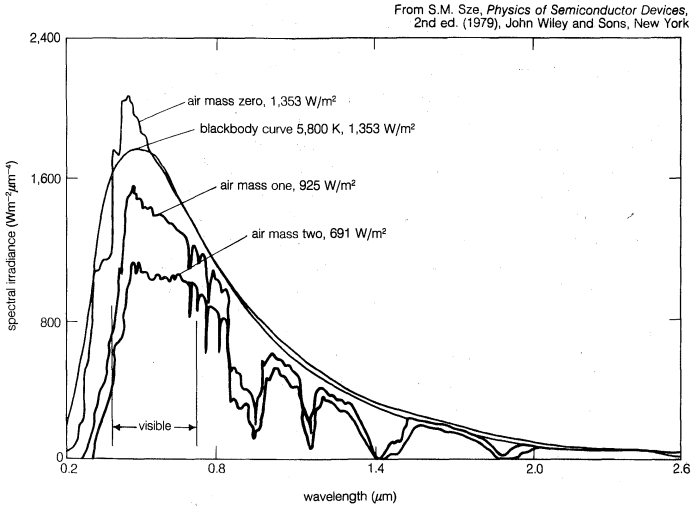


Figure 39: Solar spectral irradiance.

The efficiency of the silicon solar cell is limited to about 15 percent, because the long-wavelength emission is not absorbed near enough to the *p–n* junction for the photogenerated carriers to be collected by the *p–n* junction, and the short-wavelength emission generates carriers so near the top surface that they recombine at the surface rather than being collected by the *p–n* junction.

For satellites and space vehicles, the most important properties are conversion efficiency and reliability. Degradation due to high-energy particle radiation is an important consideration. A schematic representation of the type of solar cell developed for satellite applications is shown in Figure 40. The *p–n* junction is formed very near the front surface by diffusion. The front contact stripe and fingers must be wide enough to have low resistance but not so large as to prevent the sunlight from reaching the *p–n* junction. Thus, the conversion efficiency is decreased by that fraction of the front surface covered by the ohmic contact. The front surface also is covered by an antireflection coating. An ohmic contact covers the entire back surface.

From S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (1979); John Wiley and Sons, New York

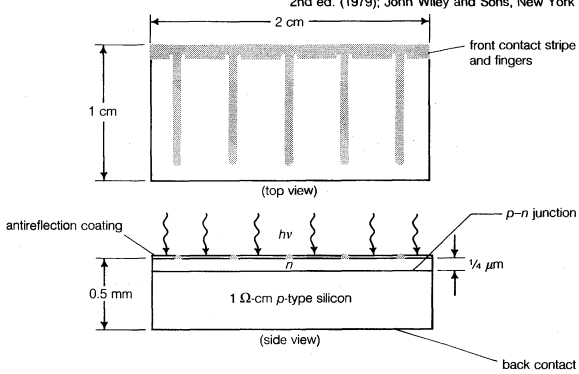


Figure 40: Silicon *p–n* junction solar cell.

A more economical fabrication technology utilizing polycrystalline silicon is shown in Figure 41. This is made by cutting wafers from cast blocks, and it measures 10 centimetres square. The treelike structure is the ohmic contact to the front surface, and the irregular-shaped regions are the silicon grains. These cells are connected in large solar panels for application on the Earth's surface to provide power at remote sites.

Use of polycrystalline silicon

Another technique involves the use of mirrors and lenses to concentrate sunlight and the utilization of a smaller cell area. With silicon cells at a concentration of 1,000 suns, the output is about the same for 1,300 cells at one sun. A gallium arsenide cell with a top layer of aluminum gallium arsenide can be useful with concentrators, and an efficiency of 23 percent close to 1,000 suns gives an output power of 10 watts.

Light-emitting diodes and semiconductor lasers. *Light-emitting diodes.* The familiar light bulb gives off light due to its temperature (incandescence). Luminescence, on the other hand, is the result of electronic excitation of a material. The light-emitting diode is a *p–n* junction in which an applied voltage yields a flow of current, and the recombination of the carriers injected across the junction results in the emission of light. (The process involved here is in effect electroluminescence.) The ratio of the number of emitted photons to the number of electrons crossing

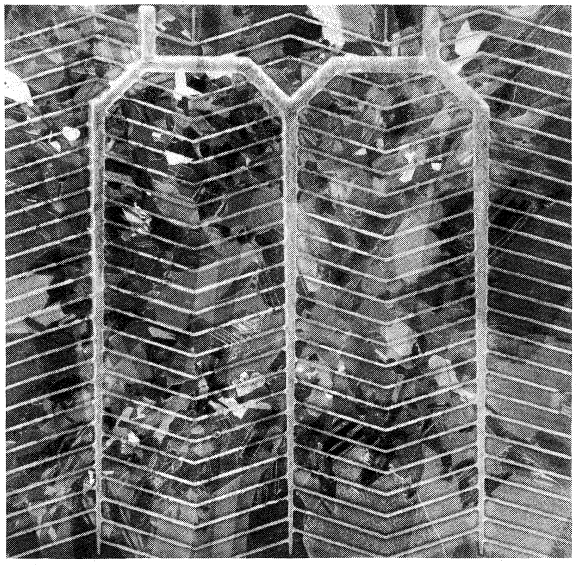


Figure 41: Four-inch-square polycrystalline silicon solar cell.

the p–n junction is the quantum efficiency. LED emission is generally in the visible part of the spectrum with wavelengths from 0.4 to 0.7 μm or in the near infrared with wavelengths between 2.0 and 0.7 μm. Visible LEDs are used as numeric displays or indicator lamps, while the infrared LEDs are employed in opto-isolators or as sources in optical communication systems. The applied voltage is near 1.5 volts. The current would depend on the application and would range from a few milliamperes to several hundred milliamperes.

Use of III-V compound semiconductors

Silicon, the most commonly used semiconductor for electronic devices and integrated circuits, is not suitable for LEDs. In silicon, the electrons in the conduction band and the holes in the valence band do not have the same momentum. Therefore, the recombination of an electron in the conduction band with a hole in the valence band necessary for the emission of a photon cannot readily occur because of the difference in momentum. However, in the III-V compound semiconductors, such as gallium arsenide (GaAs), gallium arsenide phosphide ($GaAs_{1-x}P_x$), aluminum gallium arsenide ($Al_xGa_{1-x}As$), and gallium indium arsenide phosphide ($Ga_xIn_{1-x}As_yP_{1-y}$), radiative recombination can occur with ease because electrons in the conduction band and holes in the valence band have the same momentum. The peak intensity of the emission spectrum occurs at a photon energy slightly less than the semiconductor energy gap. The radiative recombination must compete with various non-radiative recombination processes because of undesirable impurities and crystal defects, including precipitations and dislocations. Careful material processing is necessary to obtain useful LEDs.

From W.O. Groves, A.H. Herzog, and M.G. Craford, "The Effect of Nitrogen Doping on GaAsP Electroluminescent Materials," *Applied Physics Letters*, no. 19 (1970)

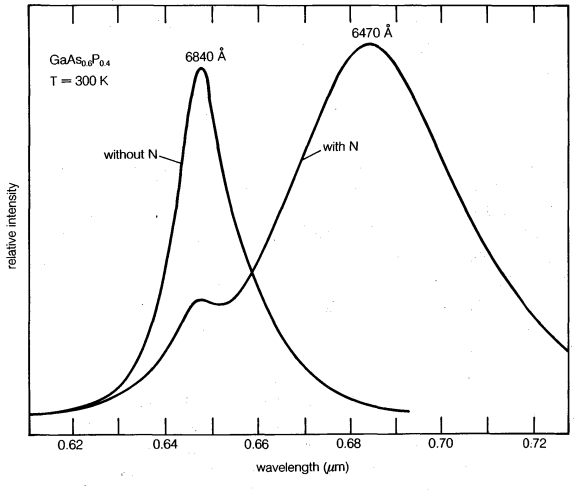


Figure 42: Emission spectra for $GaAs_{0.6}P_{0.4}$ with and without nitrogen.

In the notation used for the ternary solid solutions, the x in $Al_xGa_{1-x}As$ means that x percent of the group III elements are aluminum and $(1-x)$ percent of the group III elements are gallium. For example, with $x = 0.3$, the $Al_{0.3}Ga_{0.7}As$ would have 30 percent aluminum and 70 percent gallium. For quaternary solid solutions such as $Ga_xIn_{1-x}As_yP_{1-y}$, the x represents the percent of group III elements, while the y represents the percent of group V elements. The energy gap, and therefore the wavelength of the emitted light, changes with the composition x or x and y. The wavelength of the emitted light can be changed by varying the semiconductor composition.

The brightness of the light observed from an LED will depend on the power emitted by the LED and the relative sensitivity of the eye at the emitted wavelength. The maximum sensitivity occurs at 0.555 μm, which is in the yellow-orange and green region.

The first visible LED with extensive applications was based on $GaAs_{1-x}P_x$ grown on GaAs substrates. The composition for maximum brightness depends on both the quantum efficiency of the LED and the sensitivity of the eye. The external quantum efficiency for $GaAs_{1-x}P_x$ decreases as composition x increases, but the eye response improves as the emission moves to greater photon energy. The maximum brightness occurs near $x = 0.4$, which is an energy gap of approximately 1.9 eV. At this composition, the external quantum efficiency is about 0.2 percent.

As x increases beyond 0.45, the electron and hole no longer have the same momentum, and the energy gap goes from direct to indirect. An efficient radiative recombination centre in the indirect energy-gap region for $x > 0.45$ is the group V element nitrogen, which replaces phosphorus atoms. A recombination centre of this type is called an isoelectronic centre, and the quantum efficiency is greatly enhanced with nitrogen. Figure 42 shows that the spectrum peak moves to a longer wavelength and the emission spectrum becomes wider with the addition of nitrogen. With the isoelectronic nitrogen centre, efficient LEDs from red to orange, yellow, and green are produced.

Isoelectronic centre

From S. Gage, D. Evans, and M. Hodapp, *Optoelectronics/Fiber-Optics Applications Manual*, 2nd ed. (1977); McGraw-Hill Book Company, New York

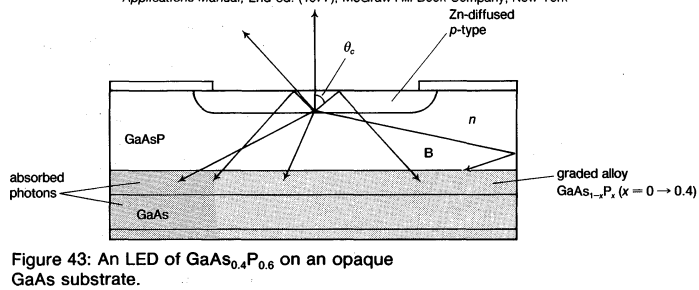


Figure 43: An LED of $GaAs_{0.4}P_{0.6}$ on an opaque GaAs substrate.

The basic configuration of a visible LED of $GaAs_{1-x}P_x$ is illustrated in Figure 43. Gallium arsenide was used as the substrate material, and the composition x varied from $x = 0$ at the substrate to 0.4. Since $GaAs_{0.4}P_{0.6}$ and GaAs do not have the same lattice constant, the composition of the layer is graded to minimize the non-radiative recombination centres that occur as a result of the lattice mismatch. The $GaAs_{1-x}P_x$ is covered with silicon dioxide and holes etched in the oxide, and then zinc is diffused into the n-type $GaAs_{1-x}P_x$ to give the p–n junction. The photons generated by radiative recombination at the p–n junction are emitted in all directions, but only a small fraction escape from the front surface. Because the refractive index of the $GaAs_{1-x}P_x$ is larger than for air, the emitted light is reflected back from the surface. At normal incidence, about 30 percent is reflected. At the critical angle represented by θ_c, all the light is internally reflected. Because the gallium arsenide substrate has a narrower energy gap than does gallium arsenide phosphide, all the light reflected to the substrate is absorbed. This internal absorption causes the internal and external quantum efficiencies to differ by a factor of approximately 10. The use of a substrate transparent to the emitted radiation, such

as gallium phosphide, and a reflecting back contact can greatly enhance the external quantum efficiency.

Any LED may be used as a source with a plastic or glass optical fibre for a short-range optical fibre transmission system over a distance of less than 100 metres. For a long-range optical-fibre transmission system, the emission properties of the source are selected to match the transmission properties of the optical fibre, and the infrared LEDs are a better match than the visible LEDs. Glass optical fibres suffer much less loss than do plastic fibres, and the glass fibres have the lowest transmission loss in the infrared at wavelengths of 1.3 and 1.55 μm. To match the transmission properties of the low-loss glass fibres, the quaternary solid solution gallium indium arsenide phosphide is grown with compositions that have the same lattice constant as the indium phosphide (InP) substrate. The composition of the gallium indium arsenide phosphide may be selected to give emission at 1.3 or 1.55 μm.

Lasers. Optical devices of this kind produce a more directional light beam and a narrow wavelength band. The term laser is an acronym derived from light amplification by stimulated emission of radiation.

Principles of operation

A photon and an electron can interact in two ways. In an absorption process, a photon with an energy slightly larger than the energy gap can interact with an electron in the valence band and raise the electron to the conduction band and create a hole in the valence band. In a stimulated emission process, a photon with an energy slightly greater than the energy gap can interact with an electron in the conduction band and cause the electron to recombine with a hole in the valence band. The recombination process results in the emission of a photon identical to the photon that caused the recombination process, and the number of photons is increased. In absorption a photon is lost, while in stimulated emission an additional photon is created. To have a laser, it is necessary to provide a structure that will make stimulated emission more probable than absorption.

A semiconductor laser requires a p–n junction that provides light emission like an LED (*i.e.,* spontaneous emission) when a voltage is applied and current flows. This radiative recombination has to be confined along the junction plane and must be reflected by parallel, partially reflecting surfaces so as to form a cavity. These parallel mirrors are readily obtained by cleaving along the natural cleavage planes of III-V compound semiconductors. The injected electrons and the light must be confined to the same region, so that they can interact to enhance the stimulated emission.

To provide the carrier and light confinement to the region of the p–n junction and obtain continuous operation at room temperature, it is necessary to use a heterojunction—*i.e.,* the junction in a single crystal between two dissimilar semiconductors. The most significant difference is the energy gap and the refractive index. A double heterostructure made with aluminum gallium arsenide and gallium arsenide is shown in Figure 44A. The left layer is n-type $Al_xGa_{1-x}As$, as the centre layer is p-type GaAs, and the right layer is p-type $Al_xGa_{1-x}As$. The centre layer of GaAs has a smaller energy gap than the two cladding layers of $Al_xGa_{1-x}As$. Typical values for x are 0.3. This gallium arsenide region with a smaller energy gap is where the light is generated due to radiative recombination of the injected carriers; it is called the active region. With an active layer of gallium arsenide, the emission wavelength is in the infrared near 0.9 μm. Other pairs of semiconductors may be used, but all require a smaller-energy-gap active region with larger-energy-gap cladding layers. Also, to prevent non-radiative recombination at the heterojunction interfaces, the active layer and the cladding layers must have the same lattice constant. It is fortunate that GaAs and $Al_xGa_{1-x}As$ closely match as to lattice constant for all compositions of x.

Figure 44B shows the energy gaps of the three regions when a voltage of approximately 1.7 is applied to the heterostructure laser. Electrons are injected from the n-layer of $Al_xGa_{1-x}As$ into the potential well of GaAs formed between the two $Al_xGa_{1-x}As$ layers. The injected carriers are confined to the narrow active layer, which typically

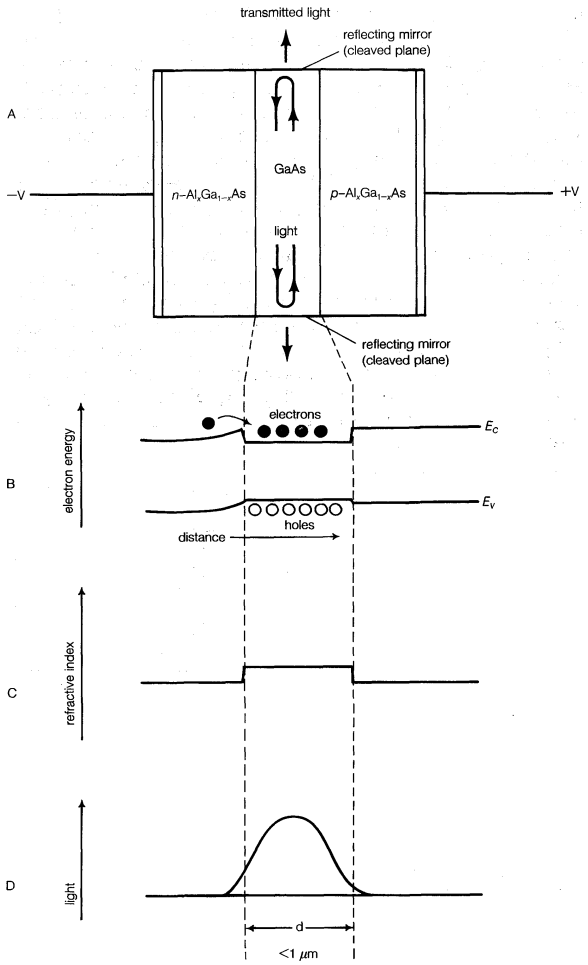


Figure 44: (A) $Al_{0.3}Ga_{0.7}As$-GaAs double heterostructure laser. (B) Energy-band diagram at forward bias. (C) Refractive-index profile. (D) Optical-field distribution.

has a thickness of between 0.1 and 0.2 μm. The refractive index profile is shown in Figure 44C. Because the $Al_xGa_{1-x}As$ layers have a smaller refractive index than the GaAs region, they form an optical wave guide that confines the light due to radiative recombination to the active layer. The light confinement is illustrated in Figure 44D. The carrier and light-confining properties of the heterojunctions permit a high density of injected electrons in the region where the light is confined, so that the photons can interact with the electrons and cause stimulated emission. The oscillation condition requires that the stimulated emission exceed the light lost through the partially reflecting facets and internal absorption.

Epitaxial growth techniques

The layers of $Al_xGa_{1-x}As$ and GaAs are grown on a GaAs substrate by a variety of epitaxial growth techniques, which include liquid-phase epitaxy, molecular-beam epitaxy, and

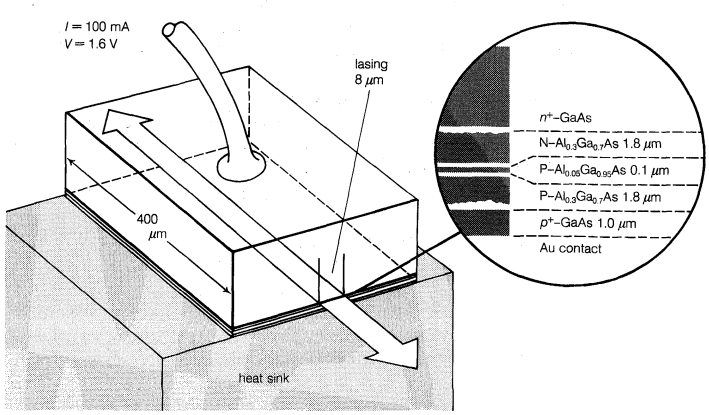


Figure 45: Stripe-geometry laser mounted on a heat sink.

metal-organic chemical-vapour deposition. These techniques permit the growth of the micrometre- and submicrometre-thick layers necessary for heterostructure lasers. Semiconductor lasers use a stripe-geometry that restricts the current along the junction plane to a width of from 1 to 15 μm. The laser chip is mounted upside down so that the heat-generating region, the p–n junction, is close to the heat sink. Figure 45 features a stripe-geometry laser mounted on a heat sink. Figure 46 shows light output as a function of the current. At low current, the light output is small, and this emission results from spontaneous emission, as in an LED. When the stimulated emission exceeds the internal losses and the light emitted from the cleaved facets, the laser threshold is reached and the light output rises rapidly with the current. Above the threshold current, most of the current flowing into the p–n junction results in laser emission, and the quantum efficiency is much higher than for an LED. The emission spectrum is also given in Figure 46.

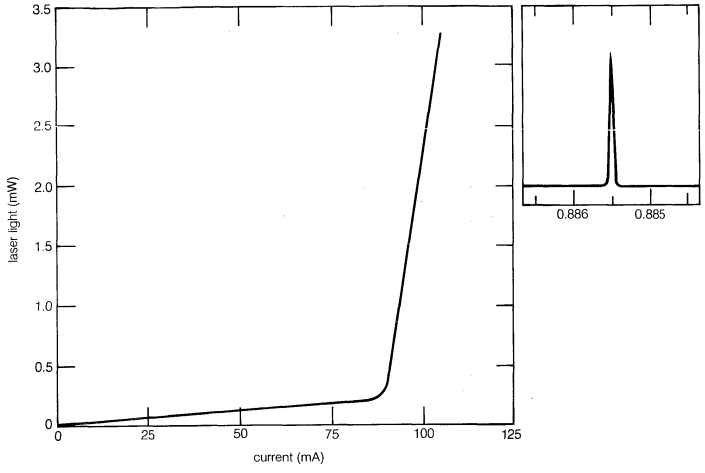


Figure 46: Light output versus the current for a heterostructure laser. The emission spectrum is shown at the upper right.

Stripe-geometry lasers typically have threshold currents near 50 mA, but lasers with threshold currents of less than 1 mA have been demonstrated. The emission may be turned off and on for digital signals by an applied pulse. The frequency response can be very fast, and operation at rates in excess of 1 GHz can readily be obtained. The best match to the transmission properties of optical fibres occurs at wavelengths of 1.3 and 1.55 μm. Heterostructure lasers for emission in this wavelength region have cladding layers made of indium phosphide and active layers consisting of gallium indium arsenide phosphide, which lattice match indium phosphide. These lasers are used as light sources in long-distance optical-fibre communication systems. Such applications include undersea cables.

Major uses of semiconductor lasers

The largest volume use of semiconductor lasers is in compact disc players. In such devices, laser light is focused on a plastic disc coated with a thin metallic film, and digital information is communicated by the presence or absence of "pits." These microscopic holes change the reflectivity of the emitted light, which is detected by light-sensitive diodes.

Optical fibres. Optical fibres are glass or plastic wave guides for transmitting visible or infrared signals. Since plastic fibres have high attenuation and are used only in limited applications, they will not be considered here. Glass fibres are frequently thinner than human hair and are generally used with LEDs or semiconductor lasers that emit in the infrared region. For wavelengths near 0.8 to 0.9 μm, gallium arsenide-aluminum gallium arsenide (GaAs-$Al_xGa_{1-x}As$) sources are used, and for those of 1.3 and 1.55 μm, indium phosphide-gallium indium arsenide phosphide (InP-$Ga_xIn_{1-x}As_yP_{1-y}$) sources are employed. As noted earlier, optical fibres consist of a glass core region that is surrounded by glass cladding. The core region has a larger refractive index than the cladding so that the light is confined to that region as it propagates along the fibre.

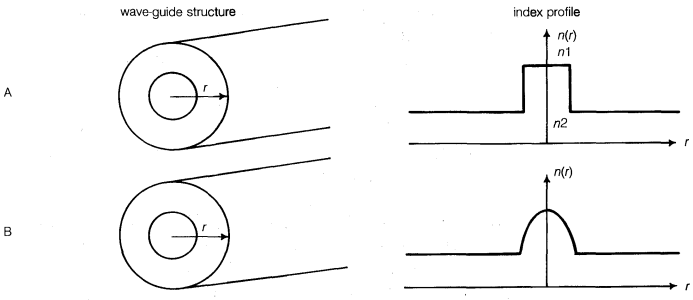


Figure 47: *Optical-fibre wave-guide structures and refractive index profiles.*
(A) Step-index fibre. (B) Graded-index fibre.

Fibre core diameters range between 1 and 100 μm, while cladding diameters are between 100 and 300 μm.

Fibres with a larger core diameter are called multimode fibres, because more than one electromagnetic-field configuration can propagate through such a fibre. A single-mode fibre has a small core diameter, and the difference in refractive index between the core and cladding is smaller than for the multimode fibre. Only one electromagnetic-field configuration propagates through a single-mode fibre. Fibres of this variety have the lowest losses and are the most widely used because they permit longer transmission distances. They have a constant refractive index in the core with a diameter between 1 and 10 μm. The index in the cladding layer decreases by approximately 0.1 to 0.3 percent. This type of fibre is called a step-index fibre and is illustrated in Figure 47A.

Single-mode optical fibres

The multimode fibres may be step-index fibres with diameters between 40 and 100 μm. The refractive index step between the core and cladding is approximately 0.8 to 3 percent. In a graded-index fibre, the core refractive index varies as a function of radial distance, as shown in Figure 47B. In such a fibre, a ray in the centre of the core travels more slowly than one near the edge, because the speed of propagation v is related to refractive index n as $v = c/n$, where c is the speed of light. The ray near the edge has a longer zigzag path than the ray in the centre. The transit times of the rays are thus equalized.

Both single-mode and multimode fibres are made of silica glass. The refractive indexes of the silica are varied with dopants such as germanium dioxide (GeO_2), phosphoric oxide (P_2O_5), and boric oxide (B_2O_3). Vapour-phase growth reactions are used to obtain the "preform" rod, which is then drawn into optical fibres. For example, a GeO_2-SiO_2 film may be deposited inside a silica tube, as shown in Figure 48. In this case, the GeO_2 increases the core refractive index. In another method, preforms for low-loss, single-mode fibres are made by first depositing a low-index borosilicate layer on the inner surface of the silica tube and then depositing a silica layer or inserting a pure fused silica rod before collapsing the preform. The preform is subsequently drawn into the optical fibre and covered with a protective polymer coating.

There are a number of factors that contribute to attenuation in an optical fibre. Rayleigh scattering is caused by microscopic variations in the refractive index of a fibre and is proportional to λ^4. This loss is the limiting loss shown by the dashed line in Figure 49 for fibre loss. Absorption by hydroxyl (OH) ions increases the absorption as shown and gives the minima in loss at 1.3 and 1.55 μm. At

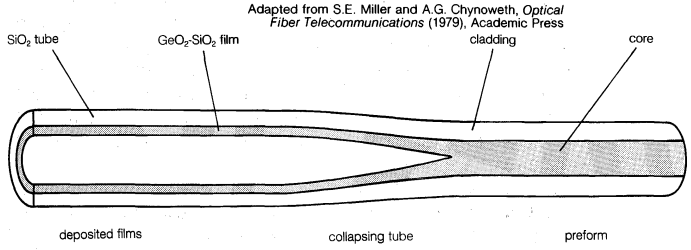


Figure 48: Preparation of the preform rod by vapour deposition. The high-index core is deposited inside the silica tube and becomes the fibre core when the tube collapses.

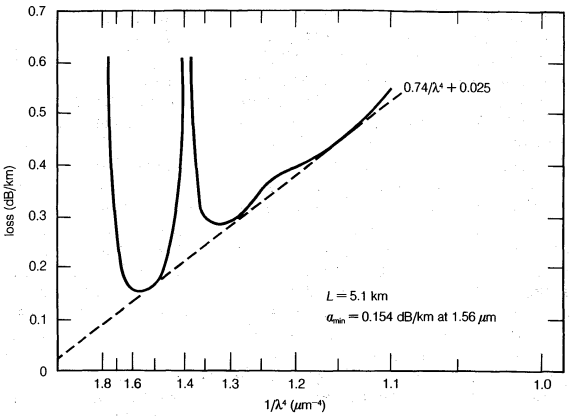


Figure 49: Low-loss, single-mode fibre attenuation plotted as a function of λ^4 (see text).

From T. Kimura, "Factors Affecting Fiber-Optic Transmission Quality," *Journal of Lightwave Technology*, 6:611, © 1988 IEEE

longer wavelengths, absorption by the atomic vibrations in the silicon-oxygen atoms rapidly increases the loss. Single-mode fibres commercially available for communications systems have losses as low as 0.2 decibel per kilometre. The low fibre loss permits increased repeater spacing and lower system cost. High-bit-rate digital systems without repeaters have been demonstrated for fibre lengths of more than 100 kilometres.

Fibre splicing techniques have been developed so that repairs can be made in the field with losses of only 0.1 to 0.3 decibel. A variety of optical connectors are used, providing both ease of use and low loss of only a few tenths of a decibel. Fibres are combined into many different kinds of cables, which can be laid both in the ground and under the sea.

From T.J. Scheffer, "Direct-Multiplexed Liquid-Crystal Displays," SID Seminar Lecture Notes, 4.2 (1987)

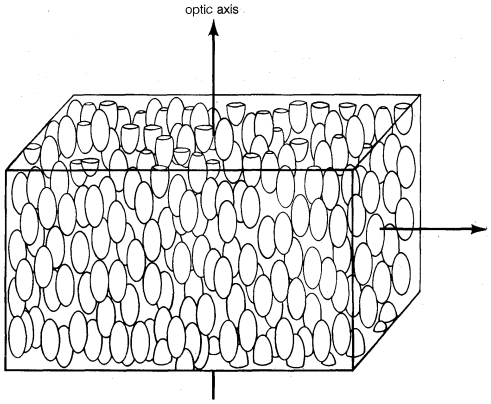


Figure 50: Molecular order in a nematic liquid crystal.

Liquid-crystal displays. Such visual display devices are commonly used as displays for digital wristwatches and pocket calculators, as indicators in automatic cameras, and as flat-panel display screens for portable lap-top computers and pocket televisions. Very little power is required to operate a liquid-crystal display (LCD) because it modifies ambient light instead of generating light. For this reason, it is the only type of electronic display that can operate for more than a year on a small battery.

Liquid crystals have both the fluidity characteristic of liquids and the collective orientation of crystals associated with solids. Their viscosity is similar to that of light machine oil. A widely used variety, the nematic liquid crystal, is composed of elongated organic molecules that align in a preferred direction called the director. This molecular order is shown in Figure 50. The nematic liquids are cloudy, white liquids, but in thin layers they are transparent because incident light will simply pass through them.

Nematic liquid crystals

For use in displays, liquid crystals must be aligned in a particular manner. To align the liquid crystals by a glass surface, an alignment layer is obtained by coating a thin polyimide film on the glass and unidirectionally rubbing the surface with a nylon brush at a precisely controlled pressure and speed. The construction of a liquid-crystal display is shown in Figure 51. The liquid crystal is sandwiched between two glass substrates that have transparent electrode strips. The transparent electrodes are made of indium-tin oxide. Spacer particles keep the plates separated uniformly.

Operation of an LCD is illustrated in Figure 52. In the field-off state (left), no voltage is applied to the electrodes. The glass surfaces at the top and bottom have been rubbed at right angles to each other. Because the liquid crystals align parallel to the rubbing direction at the glass surfaces, the crystals undergo a continuous 90° twist in between the two surfaces. This type of display is known as a twisted nematic display. Polarizing sheets are laminated on the outside of the glass parallel to the rubbing direction. The back glass also has a reflecting mirror surface. The linear

Twisted nematic display

From T.J. Scheffer, "Direct-Multiplexed Liquid-Crystal Displays," SID Seminar Lecture Notes, 4.2 (1987)

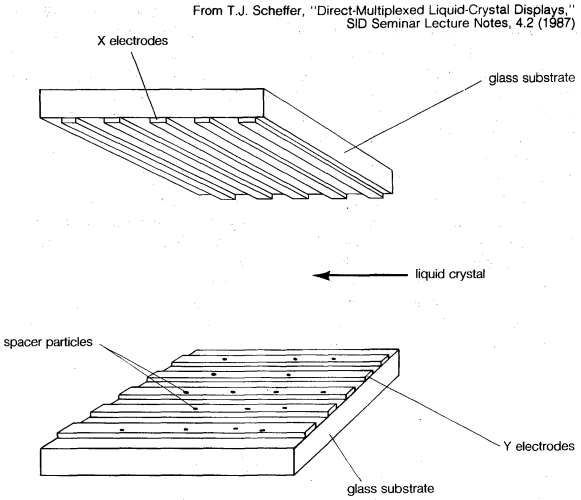


Figure 51: Construction of a liquid-crystal display.

polarized light through the upper polarizer follows the twist in the liquid crystal and passes through the polarizer at the bottom. This light is reflected by the mirror and reverses its path and emerges at the top surface. Because the light has been reflected, the area appears bright. If a small voltage is applied, the resulting electric field causes the molecules to align with the field, as shown at the right of the Figure. Now the polarized light is not rotated by the liquid crystal, and it is absorbed by the polarizer at the bottom. This area will appear dark.

The desired pattern for the display is obtained by an electrode pattern applied to the glass substrates. In a picture display, each picture element, or pixel, must be addressed.

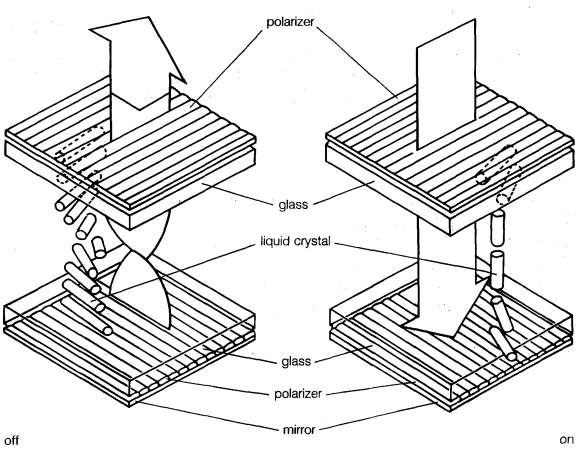


Figure 52: *Operation of a twisted nematic liquid-crystal display.* (Left) The "off" state rotates the polarized light so that it is reflected to give off a bright appearance. (Right) The "on" state with an applied voltage causes the molecules to align with the electric field, and the polarized light is not rotated and is absorbed in the bottom polarizer to give a dark appearance.

An array of 640 × 400 pixels may be involved. Multiplexing schemes are used so that 640 + 400 electrical contacts are required rather than 640 × 400 = 256,000 contacts. Advanced integrated-circuit technology has made it possible to deposit silicon directly on the glass and produce very inexpensive circuitry to drive the display. (H.C.C.)

BIBLIOGRAPHY

History of electronics: Developments in electronics are outlined in HENRY B.O. DAVIS, *Electrical and Electronic Technologies: A Chronology of Events and Inventors to 1900* (1981), *Electrical and Electronic Technologies: A Chronology of Events and Inventors from 1900 to 1940* (1983), and *Electrical and Electronic Technologies: A Chronology of Events and Inventors from 1940 to 1980* (1985); G.W.A. DUMMER, *Electronic Inventions and Discoveries: Electronics from its Earliest Beginnings to the Present Day,* 3rd rev. and expanded ed. (1983); and W.A. ATHERTON, *From Compass to Computer: A History of Electrical and Electronics Engineering* (1984).

The science of electronics: Fundamental principles and basic functions of electronics are presented in PAUL HOROWITZ and WINFIELD HILL, *The Art of Electronics* (1980); S.W. AMOS, *Principles of Transistor Circuits: Introduction to the Design of Amplifiers, Receivers, and Digital Circuits,* 6th ed. (1981), an elementary discussion of devices and circuits; J. SEYMOUR, *Electronic Devices and Components* (1981, reissued 1986); ROBERT J. MATTHYS, *Crystal Oscillator Circuits* (1983), an introductory textbook covering a wide range of oscillators; ARTHUR H. SEIDMAN (ed.), *Integrated Circuits Applications Handbook* (1983), an extensive coverage with many detailed examples; M. KUBÁT, *Power Semiconductors* (1984), a comprehensive textbook on devices for power frequency applications; DEWITT G. ONG, *Modern MOS Technology: Processes, Devices, and Designs* (1984), an introductory textbook on metal-oxide semiconductors; ROBERT BOYLESTAD and LOUIS NASHELSKY, *Electronic Devices and Circuit Theory* (1987); and ROBERT E. SIMPSON, *Introductory Electronics for Scientists and Engineers,* 2nd ed. (1987). See also JAMES T. HUMPHRIES and LESLIE P. SHEETS, *Industrial Electronics,* 2nd ed. (1986).

Reference works on electronics include DONALD G. FINK and DONALD CHRISTIANSEN (eds.), *Electronics Engineers' Handbook,* 2nd ed. (1982); FRANK JAY (ed.), *IEEE Standard Dictionary of Electrical and Electronics Terms* (1984); STAN GIBILISCO (ed.), *Encyclopedia of Electronics* (1985); *Reference Data for Engineers: Radio, Electronics, Computer, and Communications,* 7th ed. (1985); and JOHN DOUGLAS-YOUNG, *Illustrated Encyclopedic Dictionary of Electronics,* 2nd ed. (1987). See also AMERICAN RADIO RELAY LEAGUE, *The ARRL Handbook for the Radio Amateur,* 65th ed. (1988), which covers electronic and electrical principles and explains how devices work and how to apply them, assuming only modest technical knowledge.

(R.I.S.)

Electron tubes: CURTIS L. HEMENWAY, RICHARD W. HENRY, and MARTIN CAULTON, *Physical Electronics,* 2nd ed. (1967), on the fundamental physics of electron tubes; JAMES T. COLEMAN, *Microwave Devices* (1982), a general treatment of vacuum devices including fast-wave tubes; A.S. GILMOUR, JR., *Microwave Tubes* (1986), a comprehensive treatment of modern electron tubes; and SAMUEL Y. LIAO, *Microwave Electron-Tube Devices* (1988), with theoretical and experimental coverage of the basic and newer types of electron tubes.

(E.N.So.)

Semiconductor devices: Semiconductor devices and related subjects receive full coverage in the following works: R.A. SMITH, *Semiconductors,* 2nd ed. (1978), a classic text on semiconductor physics; S.M. SZE, *Physics of Semiconductor Devices,* 2nd ed. (1981), an in-depth treatment of physics and mathematical formulations of semiconductor devices, and *Semiconductor Devices: Physics and Technology* (1985), an introduction to the physical principles of semiconductor devices and their fabrication technology; W.E. BEADLE, J.C.C. TSAI, and R.D. PLUMMER (eds.), *Quick Reference Manual for Silicon Integrated Circuit Technology* (1985), an extensive collection of tables and charts for device design and fabrication; and S.M. SZE (ed.), *VLSI Technology,* 2nd ed. (1988), a complete volume on the theoretical and practical aspects of silicon-processing technology from discrete to very-large-scale integrated (VLSI) circuits.

(S.M.Sz.)

Integrated circuits: Overviews are presented in GÜNTER FRIEDRICHS and ADAM SCHAFF (eds.), *Microelectronics and Society: For Better or for Worse* (1982); T.R. REID, *The Chip: How Two Americans Invented the Microchip and Launched a Revolution* (1984, reissued 1986; U.K. title, *Microchip: The Story of a Revolution and the Men Who Made It,* 1985, reissued 1986); and in a series of articles in *Scientific American,* vol. 237, no. 3 (Sept. 1977). Integrated-circuit design is discussed in ARTHUR B. GLASER and GERALD E. SUBAK-SHARPE, *Integrated Circuit Engineering: Design, Fabrication, and Applications* (1977); PAUL R. GRAY and ROBERT G. MEYER, *Analysis and Design of Analog Integrated Circuits,* 2nd ed. (1984); TEXAS INSTRUMENTS INCORPORATED, *Linear Circuits Data Book* (1984); L.J. HERBST, *Monolithic Integrated Circuits: Techniques and Capabilities* (1985); RICHARD S. MULLER and THEODORE I. KAMINS, *Device Electronics for Integrated Circuits,* 2nd ed. (1986); and DAVID A. HODGES and HORACE G. JACKSON, *Analysis and Design of Digital Integrated Circuits,* 2nd ed. (1988). Very-large-scale integration systems are treated in CARVER MEAD and LYNN CONWAY, *Introduction to VLSI Systems* (1980); NORMAN G. EINSPRUCH (ed.), *VLSI Electronics: Microstructure Science* (1981–), with 16 vol. published by 1987; NORMAN G. EINSPRUCH, *VLSI Handbook* (1985); NEIL H.E. WESTE and KAMRAN ESHRAGHIAN, *Principles of CMOS VLSI Design: A Systems Perspective* (1985); and PAUL LOSLEBEN (ed.), *Advanced Research in VLSI* (1987), conference proceedings. See also two articles in *Proceedings of the IEEE*: WILLIAM C. HOLTON and RALPH K. CAVIN III, "A Perspective on CMOS Technology Trends," 74(12):1646–1668 (Dec. 1986); and MORTON E. JONES, WILLIAM C. HOLTON, and ROBERT STRATTON, "Semiconductors: The Key to Computational Plenty," 70(12):1380–1409 (Dec. 1982). Fabrication of integrated circuits is presented in IVOR BRODIE and JULIUS J. MURAY, *The Physics of Microfabrication* (1982), a moderately detailed review; SORAB K. GHANDI, *VLSI Fabrication Principles* (1983), a review on silicon and gallium arsenide fabrication technology; and DAVID J. ELLIOTT, *Microlithography: Process Technology for IC Fabrication* (1986), a monograph on materials, processes, and equipment that are used in microfabrication.

(W.C.Ho.)

Optoelectronics: Several optoelectronic devices are discussed in S.M. SZE, *Physics of Semiconductor Devices,* 2nd ed. (1979), especially ch. 12–14. Works on specific devices include, on photodetectors, G.E. STILLMAN, V.M. ROBBINS, and N. TABATABAIE, "III-V Compound Semiconductor Devices: Optical Detectors," *IEEE Transactions on Electron Devices,* ED-31(11): 1643–1655 (Nov. 1984); and R.K. WILLARDSON and ALBERT C. BEER (eds.), *Infrared Detectors II,* vol. 12 of *Semiconductors and Semimetals* (1977); on solar cells, HAROLD J. HOVEL (ed.), *Solar Cells,* vol. 11 of *Semiconductors and Semimetals,* ed. by R.K. WILLARDSON and A.C. BEER (1975); and INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS, *The Conference Record of the Nineteenth IEEE Photovoltaic Specialists Conference—1987* (1987); on light-emitting diodes, H.C. CASEY, JR., and F.A. TRUMBORE, "Single Crystal Electroluminescent Materials," *Materials Science and Engineering,* 6(2):69–109 (Aug. 1970); and STAN GAGE *et al., Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed. (1981); on lasers, G.P. AGRAWAL and N.K. DUTTA, *Long-Wavelength Semiconductor Lasers* (1986); H.C. CASEY, JR., and M.B. PANISH, *Heterostructure Lasers,* 2 vol. (1978); and W.T. TSANG (ed.), *Lightwave Communications Technology: Part C: Semiconductor Injection Lasers II, Light Emitting Diodes,* vol. 22 in *Semiconductors and Semimetals,* ed. by R.K. WILLARDSON and ALBERT C. BEER (1985); on optical fibres, TATSUYA KIMURA, "Factors Affecting Fiber-Optic Transmission Quality," *Journal of Lightwave Technology,* 6(5):611–619 (May 1988); JOHN GOWAR, *Optical Communication Systems* (1984); and STEWART E. MILLER and ALAN G. CHYNOWETH, *Optical Fiber Telecommunications* (1979); and YASUHARU SUEMATSU and KEN-ICHI IGA, *Introduction to Optical Fiber Communications,* trans. from Japanese (1982); and on liquid-crystal displays, F.J. KAHN, "The Molecular Physics of Liquid-Crystal Devices," *Physics Today,* 35(5):66–74 (May 1982); TERRY J. SCHEFFER, "Direct Multiplexed Liquid-Crystal Displays," Society for Information Display *Seminar Lecture Notes,* 1:4/1–4/34 (1987); W.H. DE JEU, *Physical Properties of Liquid Crystalline Materials* (1980); S. CHANDRESEKHAR, *Liquid Crystals* (1977, reprinted 1980); and L. LIEBERT (ed.), *Liquid Crystals* (1978).

(H.C.C.)

Elizabeth I of England

Elizabeth I was queen of England from 1558 to 1603. Though her small kingdom was threatened by grave internal divisions, Elizabeth's blend of shrewdness, courage, and majestic self-display inspired ardent expressions of loyalty and helped unify the nation against foreign enemies. The adulation bestowed upon her both in her lifetime and in the ensuing centuries was not altogether a spontaneous effusion; it was the result of a carefully crafted, brilliantly executed campaign in which the queen fashioned herself as the glittering symbol of the nation's destiny. This political symbolism, common to monarchies, had more substance than usual, for the queen was by no means a mere figurehead. While she did not wield the absolute power of which Renaissance rulers dreamed, she tenaciously upheld her authority to make critical decisions and to set the central policies of both state and church. The latter half of the 16th century in England is justly called the Elizabethan era: rarely has the collective life of a whole age been given so distinctively personal a stamp.

From the Woburn Abbey Collection, by kind permission of His Grace, the Duke of Bedford

Elizabeth I, the Armada portrait by Gower (d. 1596). In Woburn Abbey, Bedfordshire.

Childhood. Elizabeth's early years were not auspicious. She was born at Greenwich Palace on Sept. 7, 1533, the daughter of the Tudor king Henry VIII and his second wife, Anne Boleyn. Henry had defied the pope and broken England from the authority of the Roman Catholic church in order to dissolve his marriage with his first wife, Catherine of Aragon, who had borne him a daughter, Mary. Since the king ardently hoped that Anne Boleyn would give birth to the male heir regarded as the key to stable dynastic succession, the birth of a second daughter was a bitter disappointment that dangerously weakened the new queen's position. Before Elizabeth reached her third birthday, her father had her mother beheaded on charges of adultery and treason. Moreover, at Henry's instigation, an act of Parliament declared his marriage with Anne Boleyn invalid from the beginning, thus making their daughter Elizabeth illegitimate, as Roman Catholics had all along claimed her to be. (Apparently the king was undeterred by the logical inconsistency of simultaneously invalidating the marriage and accusing his wife of adultery.) The emotional impact of these events on the little girl, who had been brought up from infancy in a separate household at Hatfield, is not known; presumably no one thought it worth recording. What was noted was her precocious seriousness; at six years old, it was admiringly observed, she had as much gravity as if she had been 40.

Loss of mother

When in 1537 Henry's third wife, Jane Seymour, gave birth to a son, Edward, Elizabeth receded still further into relative obscurity, but she was not neglected. Despite his capacity for monstrous cruelty, Henry VIII treated all his children with what contemporaries regarded as affection; Elizabeth was present at ceremonial occasions and was declared third in line to the throne. She spent much of the time with her half brother Edward and, from her 10th year onward, profited from the loving attention of her stepmother, Catherine Parr, the king's sixth and last wife. Under a series of distinguished tutors, of whom the best known is the Cambridge humanist Roger Ascham, Elizabeth received the rigorous education normally reserved for male heirs, consisting of a course of studies centring on classical languages, history, rhetoric, and moral philosophy. "Her mind has no womanly weakness," Ascham wrote with the unselfconscious sexism of the age, "her perseverance is equal to that of a man, and her memory long keeps what it quickly picks up." In addition to Greek and Latin, she became fluent in French and Italian, attainments of which she was proud and which were in later years to serve her well in the conduct of diplomacy. Thus steeped in the secular learning of the Renaissance, the quick-witted and intellectually serious princess also studied theology, imbibing the tenets of English Protestantism in its formative period. Her association with the Reformation is critically important, for it shaped the future course of the nation, but it does not appear to have been a personal passion: observers noted the young princess's fascination more with languages than with religious dogma.

Education

Position under Edward VI and Mary. With her father's death in 1547 and the accession to the throne of her frail 10-year-old brother Edward, Elizabeth's life took a perilous turn. Her guardian, the dowager queen Catherine Parr, almost immediately married Thomas Seymour, the lord high admiral. Handsome, ambitious, and discontented, Seymour began to scheme against his powerful older brother, Edward Seymour, protector of the realm during Edward VI's minority. In January 1549, shortly after the death of Catherine Parr, Thomas Seymour was arrested for treason and accused of plotting to marry Elizabeth in order to rule the kingdom. Repeated interrogations of Elizabeth and her servants led to the charge that even when his wife was alive Seymour had on several occasions behaved in a flirtatious and overly familiar manner toward the young princess. Under humiliating close questioning and in some danger, Elizabeth was extraordinarily circumspect and poised. When she was told that Seymour had been beheaded, she betrayed no emotion.

The need for circumspection, self-control, and political acumen became even greater after the death of the Protestant Edward in 1553 and the accession of Elizabeth's older half sister Mary, a religious zealot set on returning England, by force if necessary, to the Roman Catholic faith. This attempt, along with her unpopular marriage to the ardently Catholic king Philip II of Spain, aroused bitter Protestant opposition. In a charged atmosphere of treasonous rebellion and inquisitorial repression, Elizabeth's life was in grave danger. For though, as her sister demanded, she conformed outwardly to official Catholic observance, she inevitably became the focus and the obvious beneficiary of plots to overthrow the government and restore Protestantism. Arrested and sent to the Tower of London after Sir Thomas Wyatt's rebellion in January 1554, Elizabeth narrowly escaped her mother's fate. Two months later, after extensive interrogation and spying had revealed no conclusive evidence of treason on her part, she was released from the Tower and placed in close custody for a year at Woodstock. The difficulty of her situation eased somewhat, though she was never far from suspicious scrutiny. Throughout the unhappy years of Mary's childless reign, with its burning of Protestants and its military disasters, Elizabeth had continually to protest her innocence, affirm her unwavering loyalty, and proclaim

Sent to the Tower

her pious abhorrence of heresy. It was a sustained lesson in survival through self-discipline and the tactful manipulation of appearances.

Many Protestants and Roman Catholics alike assumed that her self-presentation was deceptive, but Elizabeth managed to keep her inward convictions to herself, and in religion as in much else they have remained something of a mystery. There is with Elizabeth a continual gap between a dazzling surface and an interior that she kept carefully concealed. Observers were repeatedly tantalized with what they thought was a glimpse of the interior, only to find that they had been shown another facet of the surface. Everything in Elizabeth's early life taught her to pay careful attention to how she represented herself and how she was represented by others. She learned her lesson well.

Accession. At the death of Mary on Nov. 17, 1558, Elizabeth came to the throne amid bells, bonfires, patriotic demonstrations, and other signs of public jubilation. Her entry into London and the great coronation procession that followed were masterpieces of political courtship. "If ever any person," wrote one enthusiastic observer, "had either the gift or the style to win the hearts of people, it was this Queen, and if ever she did express the same it was at that present, in coupling mildness with majesty as she did, and in stately stooping to the meanest sort." Elizabeth's smallest gestures were scrutinized for signs of the policies and tone of the new regime: When an old man in the crowd turned his back on the new queen and wept, Elizabeth exclaimed confidently that he did so out of gladness; when a girl in an allegorical pageant presented her with a Bible in English translation—banned under Mary's reign—Elizabeth kissed the book, held it up reverently, and then laid it on her breast; and when the abbot and monks of Westminster Abbey came to greet her in broad daylight with candles in their hands, she briskly dismissed them with the words "Away with those torches! we can see well enough." Spectators were thus assured that under Elizabeth England had returned, cautiously but decisively, to the Reformation.

The first weeks of her reign were not entirely given over to symbolic gestures and public ceremonial. The queen began at once to form her government and issue proclamations. She reduced the size of the Privy Council, in part to purge some of its Catholic members and in part to make it more efficient as an advisory body; she began a restructuring of the enormous royal household; she carefully balanced the need for substantial administrative and judicial continuity with the desire for change; and she assembled a core of experienced and trustworthy advisers, including William Cecil, Nicholas Bacon, Francis Walsingham, and Nicholas Throckmorton. Chief among these was Cecil (afterward Lord Burghley), whom Elizabeth appointed her principal secretary of state on the morning of her accession and who was to serve her (first in this capacity and after 1571 as lord treasurer) with remarkable sagacity and skill for 40 years.

The woman ruler in a patriarchal world. In the last year of Mary's reign, the Scottish Calvinist preacher John Knox wrote in his *The First Blast of the Trumpet Against the Monstruous Regiment of Women* that "God hath revealed to some in this our age that it is more than a monster in nature that a woman should reign and bear empire above man." With the accession of the Protestant Elizabeth, Knox's trumpet was quickly muted, but there remained a widespread conviction, reinforced by both custom and teaching, that, while men were naturally endowed with authority, women were temperamentally, intellectually, and morally unfit to govern. Men saw themselves as rational beings; they saw women as creatures likely to be dominated by impulse and passion. Gentlemen were trained in eloquence and the arts of war; gentlewomen were urged to keep silent and attend to their needlework. In men of the upper classes a will to dominate was admired or at least assumed; in women it was viewed as dangerous or grotesque.

A woman on the throne

Apologists for the queen countered that there had always been significant exceptions, such as the biblical Deborah, the prophetess who had judged Israel. Crown lawyers, moreover, elaborated a mystical legal theory known as "the king's two bodies." When she ascended the throne, according to this theory, the queen's whole being was profoundly altered: her mortal "body natural" was wedded to an immortal "body politic." "I am but one body, naturally considered," Elizabeth declared in her accession speech, "though by [God's] permission a Body Politic to govern." Her body of flesh was subject to the imperfections of all human beings (including those specific to womankind), but the body politic was timeless and perfect. Hence in theory the queen's gender was no threat to the stability and glory of the nation.

Elizabeth made it immediately clear that she intended to rule in more than name only and that she would not subordinate her judgment to that of any one individual or faction. Since her sister's reign did not provide a satisfactory model for female authority, Elizabeth had to improvise a new model, one that would overcome the considerable cultural liability of her sex. Moreover, quite apart from this liability, any English ruler's power to compel obedience had its limits. The monarch was at the pinnacle of the state, but that state was relatively impoverished and weak, without a standing army, an efficient police force, or a highly developed, effective bureaucracy. To obtain sufficient revenue to govern, the crown had to request subsidies and taxes from a potentially fractious and recalcitrant Parliament. Under these difficult circumstances, Elizabeth developed a strategy of rule that blended imperious command with an extravagant, histrionic cult of love.

The Virgin Queen

The cult of Elizabeth as the Virgin Queen wedded to her kingdom was a gradual creation that unfolded over many years, but its roots may be glimpsed at least as early as 1555. At that time, according to a report that reached the French court, Queen Mary had proposed to marry her sister to the staunchly Catholic duke of Savoy; the usually cautious and impassive Elizabeth burst into tears, declaring that she had no wish for any husband. Other matches were proposed and summarily rejected. But in this vulnerable period of her life there were obvious reasons for Elizabeth to bide her time and keep her options open. No one—not even the princess herself—need have taken very seriously her professed desire to remain single. When she became queen, speculation about a suitable match immediately intensified, and the available options became a matter of grave national concern. Beyond the general conviction that the proper role for a woman was that of a wife, the dynastic and diplomatic stakes in the projected royal marriage were extremely high. If Elizabeth died childless, the Tudor line would come to an end. The nearest heir was Mary, Queen of Scots, the granddaughter of Henry VIII's sister Margaret. Mary, a Catholic whose claim was supported by France and other powerful Catholic states, was regarded by Protestants as a nightmarish threat that could best be averted if Elizabeth produced a Protestant heir.

The queen's marriage was critical not only for the question of succession but also for the tangled web of international diplomacy. England, isolated and militarily weak, was sorely in need of the major alliances that an advantageous marriage could forge. Important suitors eagerly came forward: Philip II of Spain, who hoped to renew the link between Catholic Spain and England; Archduke Charles of Austria; Erik XIV, king of Sweden; Henry, Duke d'Anjou and later king of France; François, Duke d' Alençon; and others. Many scholars think it unlikely that Elizabeth ever seriously intended to marry any of these aspirants to her hand, for the dangers always outweighed the possible benefits, but she skillfully played one off against another and kept the marriage negotiations going for months, even years, at one moment seeming on the brink of acceptance, at the next veering away toward vows of perpetual virginity. "She is a Princess," the French ambassador remarked, "who can act any part she pleases."

English suitors

Elizabeth was courted by English suitors as well, most assiduously by her principal favourite, Robert Dudley, Earl of Leicester. As master of the horse and a member of the Privy Council, Leicester was constantly in attendance on the queen, who displayed toward him all the signs of an ardent romantic attachment. When in September 1560 Leicester's wife, Amy Robsart, died in a suspicious fall, the favourite seemed poised to marry his royal mis-

tress—so at least widespread rumours had it—but, though the queen's behaviour toward him continued to generate scandalous gossip, the decisive step was never taken. Elizabeth's resistance to a marriage she herself seemed to desire may have been politically motivated, for Leicester had many enemies at court and an unsavory reputation in the country at large. But in October 1562 the queen nearly died of smallpox, and, faced with the real possibility of a contested succession and a civil war, even rival factions were likely to have countenanced the marriage.

Probably at the core of Elizabeth's decision to remain single was an unwillingness to compromise her power. Sir Robert Naunton recorded that the queen once said angrily to Leicester, when he tried to insist upon a favour, "I will have here but one mistress and no master." To her ministers she was steadfastly loyal, encouraging their frank counsel and weighing their advice, but she did not cede ultimate authority even to the most trusted. Though she patiently received petitions and listened to anxious advice, she zealously retained her power to make the final decision in all crucial affairs of state. Unsolicited advice could at times be dangerous: when in 1579 a pamphlet was published vehemently denouncing the queen's proposed marriage to the Catholic Duke d'Alençon, its author John Stubbs and his publisher William Page were arrested and had their right hands chopped off.

Elizabeth's performances—her displays of infatuation, her apparent inclination to marry the suitor of the moment—often convinced even close advisers, so that the level of intrigue and anxiety, always high in royal courts, often rose to a feverish pitch. Far from trying to allay the anxiety, the queen seemed to augment and use it, for she was skilled at manipulating factions. This skill extended beyond marriage negotiations and became one of the hallmarks of her regime. A powerful nobleman would be led to believe that he possessed unique influence over the queen, only to discover that a hated rival had been led to a comparable belief. A golden shower of royal favour—apparent intimacies, public honours, the bestowal of such valuable perquisites as land grants and monopolies—would give way to royal aloofness or, still worse, to royal anger. The queen's anger was particularly aroused by challenges to what she regarded as her prerogative (whose scope she cannily left undefined) and indeed by any unwelcome signs of independence. The courtly atmosphere of vivacity, wit, and romance would then suddenly chill, and the queen's behaviour, as her godson Sir John Harington put it, "left no doubtings whose daughter she was." This identification of Elizabeth with her father, and particularly with his capacity for wrath, is something that the queen herself—who never made mention of her mother—periodically invoked.

Relations with Parliament

A similar blend of charm and imperiousness characterized the queen's relations with Parliament, on which she had to depend for revenue. Many sessions of Parliament, particularly in the early years of her rule, were more than cooperative with the queen; they had the rhetorical air of celebrations. But under the strain of the marriage-and-succession question, the celebratory tone, which masked serious policy differences, began over the years to wear thin, and the sessions involved complicated, often acrimonious negotiations between crown and commons. More radical members of Parliament wanted to include in debate broad areas of public policy; the queen's spokesmen struggled to restrict free discussion to government bills. Elizabeth had a rare gift for combining calculated displays of intransigence with equally calculated displays of graciousness and, on rare occasions, a prudent willingness to concede. Whenever possible, she transformed the language of politics into the language of love, likening herself to the spouse or the mother of her kingdom. Characteristic of this rhetorical strategy was her famous "Golden Speech" of 1601, when, in the face of bitter parliamentary opposition to royal monopolies, she promised reforms:

> I do assure you, there is no prince that loveth his subjects better, or whose love can countervail our love. There is no jewel, be it of never so rich a price, which I set before this jewel; I mean, your love: for I do more esteem of it, than of any treasure or riches.

A discourse of rights or interests thus became a discourse of mutual gratitude, obligation, and love. "We all loved her," Harington wrote with just a trace of irony, "for she said she loved us." In her dealings with parliamentary delegations, as with suitors and courtiers, the queen contrived to turn her gender from a serious liability into a distinct advantage.

Religious questions and the fate of Mary, Queen of Scots. Elizabeth restored England to Protestantism. The Act of Supremacy, passed by Parliament and approved in 1559, revived the antipapal statutes of Henry VIII and declared the queen supreme governor of the church, while the Act of Uniformity established a slightly revised version of the second Edwardian prayer book as the official order of worship. Elizabeth's government moved cautiously but steadily to transfer these structural and liturgical reforms from the statute books to the local parishes throughout the kingdom. Priests, temporal officers, and men proceeding to university degrees were required to swear an oath to the royal supremacy or lose their positions; absence from Sunday church service was punishable by a fine; royal commissioners sought to ensure doctrinal and liturgical conformity. Many of the nobles and gentry, along with a majority of the common people, remained loyal to the old faith, but all the key positions in the government and church were held by Protestants who employed patronage, pressure, and propaganda, as well as threats, to secure an outward observance of the religious settlement.

Militant Protestants

But to militant Protestants, including exiles from the reign of Queen Mary newly returned to England from Calvinist Geneva and other centres of continental reform, these measures seemed hopelessly pusillanimous and inadequate. They pressed for a drastic reform of the church hierarchy and church courts, a purging of residual Catholic elements in the prayer book and ritual, and a vigorous searching out and persecution of recusants. Each of these demands was repugnant to the queen. She felt that the reforms had gone far enough and that any further agitation would provoke public disorder, a dangerous itch for novelty, and an erosion of loyalty to established authority. Elizabeth, moreover, had no interest in probing the inward convictions of her subjects; provided that she could obtain public uniformity and obedience, she was willing to let the private beliefs of the heart remain hidden. This policy was consistent with her own survival strategy, her deep conservatism, and her personal dislike of evangelical fervour. When in 1576 the archbishop of Canterbury, Edmund Grindal, refused the queen's orders to suppress certain reformist educational exercises, called "propheseyings," Grindal was suspended from his functions and never restored to them. Upon Grindal's death, Elizabeth appointed a successor, Archbishop Whitgift, who vigorously pursued her policy of an authoritarian ecclesiastical regime and a relentless hostility to Puritan reformers.

The Catholic opposition

If Elizabeth's religious settlement was threatened by Protestant dissidents, it was equally threatened by the recalcitrance and opposition of English Catholics. At first this opposition seemed relatively passive, but a series of crises in the late 1560s and early '70s disclosed its potential for serious, even fatal, menace. In 1569 a rebellion of feudal aristocrats and their followers in the staunchly Catholic north of England was put down by savage military force; while in 1571 the queen's informers and spies uncovered an international conspiracy against her life, known as the Ridolfi Plot. Both threats were linked at least indirectly to Mary, Queen of Scots, who had been driven from her own kingdom in 1568 and had taken refuge in England. The presence, more prisoner than guest, of the woman whom the Roman Catholic church regarded as the rightful queen of England posed a serious political and diplomatic problem for Elizabeth, a problem greatly exacerbated by Mary's restless ambition and penchant for conspiracy. Elizabeth judged that it was too dangerous to let Mary leave the country, but at the same time she firmly rejected the advice of Parliament and many of her councillors that Mary should be executed. So a captive, at once ominous, malevolent, and pathetic, Mary remained.

The alarming increase in religious tension, political intrigue, and violence was not only an internal, English

concern. In 1570 Pope Pius V excommunicated Elizabeth and absolved her subjects from any oath of allegiance that they might have taken to her. The immediate effect was to make life more difficult for English Catholics, who were the objects of a suspicion that greatly intensified in 1572 after word reached England of the St. Bartholomew's Day massacre of Protestants (Huguenots) in France. Tension and official persecution of recusants increased in the wake of the daring clandestine missionary activities of English Jesuits, trained on the Continent and smuggled back to England. Elizabeth was under great pressure to become more involved in the continental struggle between Roman Catholics and Protestants, in particular to aid the rebels fighting the Spanish armies in the Netherlands. But she was very reluctant to become involved, in part because she detested rebellion, even rebellion undertaken in the name of Protestantism, and in part because she detested expenditures. Eventually, after vacillations that drove her councillors to despair, she agreed first to provide some limited funds and then, in 1585, to send a small expeditionary force to the Netherlands.

Fears of an assassination attempt against Elizabeth increased after Pope Gregory XIII proclaimed in 1580 that it would be no sin to rid the world of such a miserable heretic. In 1584 Europe's other major Protestant leader, William of Orange, was assassinated. Elizabeth herself showed few signs of concern—throughout her life she was a person of remarkable personal courage—but the anxiety of the ruling elite was intense. In an ugly atmosphere of intrigue, torture and execution of Jesuits, and rumours of foreign plots to kill the queen and invade England, Elizabeth's Privy Council drew up a Bond of Association, pledging its signers, in the event of an attempt on Elizabeth's life, to kill not only the assassins but also the claimant to the throne in whose interest the attempt had been made. The Association was clearly aimed at Mary, whom government spies, under the direction of Sir Francis Walsingham, had by this time discovered to be thoroughly implicated in plots against the queen's life. When Walsingham's men in 1586 uncovered the Babington Plot, another conspiracy to murder Elizabeth, the wretched Queen of Scots, her secret correspondence intercepted and her involvement clearly proved, was doomed. Mary was tried and sentenced to death. Parliament petitioned that the sentence be carried out without delay. For three months the queen hesitated and then with every sign of extreme reluctance signed the death warrant. When the news was brought to her that on Feb. 8, 1587, Mary had been beheaded, Elizabeth responded with an impressive show of grief and rage. She had not, she wrote to Mary's son, James VI of Scotland, ever intended that the execution actually take place, and she imprisoned the man who had delivered the signed warrant. It is impossible to know how many people believed Elizabeth's professions of grief; Catholics on the Continent wrote bitter denunciations of the queen, while Protestants throughout the kingdom enthusiastically celebrated the death of a woman they had feared and hated.

The Babington Plot

For years Elizabeth had cannily played a complex diplomatic game with the rival interests of France and Spain, a game comparable to her domestic manipulation of rival factions. State-sanctioned privateering raids, led by Sir Francis Drake and others, on Spanish shipping and ports alternated with conciliatory gestures and peace talks. But by the mid-1580s it became increasingly clear that England could not avoid a direct military confrontation with Spain. Word reached London that the Spanish king, Philip II, had begun to assemble an enormous fleet that would sail to the Netherlands, join forces with a waiting Spanish army led by the duke of Parma, and then proceed to an invasion and conquest of Protestant England. Always reluctant to spend money, the queen had nonetheless authorized sufficient funds during her reign to maintain a fleet of maneuverable, well-armed fighting ships, to which could be added other vessels from the merchant fleet. When in July 1588 the Invincible Armada reached English waters, the queen's ships, in one of the most famous naval encounters of history, defeated the enemy fleet, which then in an attempt to return to Spain was all but destroyed by terrible storms.

The Invincible Armada

At the moment when the Spanish invasion was imminently expected, Elizabeth resolved to review in person a detachment of soldiers assembled at Tilbury. Dressed in a white gown and a silver breastplate, she rode through the camp and proceeded to deliver a celebrated speech. Some of her councillors, she said, had cautioned her against appearing before a large, armed crowd, but she did not and would not distrust her faithful and loving people. Nor was she afraid of Parma's army: "I know I have the body of a weak and feeble woman," Elizabeth declared, "but I have the heart and stomach of a king, and of a king of England too." She then promised, "in the word of a Prince," richly to reward her loyal troops, a promise that she characteristically proved reluctant to keep. The scene exemplifies many of the queen's qualities: her courage, her histrionic command of grand public occasions, her rhetorical blending of magniloquence and the language of love, her strategic identification with martial virtues considered male, and even her princely parsimony.

The queen's image. Elizabeth's parsimony did not extend to personal adornments. She possessed a vast repertory of fantastically elaborate dresses and rich jewels. Her passion for dress was bound up with political calculation and an acute self-consciousness about her image. She tried to control the royal portraits that circulated widely in England and abroad, and her appearances in public were dazzling displays of wealth and magnificence. Throughout her reign she moved restlessly from one of her palaces to another—Whitehall, Nonsuch, Greenwich, Windsor, Richmond, Hampton Court, and Oatlands—and availed herself of the hospitality of her wealthy subjects. On her journeys, known as royal progresses, she wooed her people and was received with lavish entertainments. Artists, including poets like Edmund Spenser and painters like Nicholas Hilliard, celebrated her in a variety of mythological guises—as Diana, the chaste goddess of the moon; Astraea, the goddess of justice; Gloriana, the queen of the fairies—and Elizabeth, in addition to adopting these fanciful roles, appropriated to herself some of the veneration that pious Englishmen had directed to the Virgin Mary.

"She imagined," wrote Francis Bacon a few years after the queen's death, "that the people, who are much influenced by externals, would be diverted by the glitter of her jewels, from noticing the decay of her personal attractions." Bacon's cynicism reflects the darkening tone of the last decade of Elizabeth's reign, when her control over her country's political, religious, and economic forces and over her representation of herself began to show severe strains. Bad harvests, persistent inflation, and unemployment caused hardship and a loss of public morale. Charges of corruption and greed led to widespread popular hatred of many of the queen's favourites to whom she had given lucrative and much-resented monopolies. A series of disastrous military attempts to subjugate the Irish culminated in a crisis of authority with her last great favourite, Robert Devereux, the proud Earl of Essex, who had undertaken to defeat rebel forces led by Hugh O'Neill, Earl of Tyrone. Essex returned from Ireland against the queen's orders, insulted her in her presence, and then made a desperate, foolhardy attempt to raise an insurrection. He was tried for treason and executed on Feb. 25, 1601.

The last decade

Elizabeth continued to make brilliant speeches, to exercise her authority, and to receive the extravagant compliments of her admirers, but she was, as Sir Walter Raleigh remarked, "a lady surprised by time," and her long reign was drawing to a close. She suffered from bouts of melancholy and ill health and showed signs of increasing debility. Her more astute advisers—among them Lord Burghley's son, Sir Robert Cecil, who had succeeded his father as her principal counselor—secretly entered into correspondence with the likeliest claimant to the throne, James VI of Scotland. On March 24, 1603, having reportedly indicated James as her successor, Elizabeth died quietly. The nation enthusiastically welcomed its new king. But in a very few years the English began to express nostalgia for the rule of "Good Queen Bess." Long before her death she had transformed herself into a powerful image of female authority, regal magnificence, and national pride, and that image has endured to the present.

BIBLIOGRAPHY

Writings by Elizabeth: Some of Elizabeth's private letters appear in *The Letters of Queen Elizabeth,* ed. by G.B. HARRISON (1935, reprinted 1981); others are included in *The Girlhood of Queen Elizabeth: A Narrative in Contemporary Letters,* ed. by FRANK A. MUMBY (1909). Both of these volumes, however, include letters whose authenticity is doubtful. Elizabeth's translations of classical verse by Boethius, Plutarch, and Horace are published in *Queen Elizabeth's Englishings . . . ,* ed. by CAROLINE PEMBERTON (1899, reprinted 1975); and her poetry appears in *The Poems of Queen Elizabeth I,* ed. by LEICESTER BRADNER (1964). A brief sampling of her speeches may be found in *The Public Speaking of Queen Elizabeth: Selections from the Official Addresses,* ed. by GEORGE P. RICE, JR. (1951, reissued 1966); a more complete selection is available in J.E. NEALE, *Elizabeth I and Her Parliaments,* 2 vol. (1953–57, reissued 1966), which reprints complete transcripts of the queen's known addresses to Parliament. The speeches she made while on royal progresses are included in JOHN NICHOLS, *The Progresses and Public Processions of Queen Elizabeth,* new ed., 3 vol. (1823, reprinted 1966).

Biographies: The standard biography of Elizabeth remains J.E. NEALE, *Queen Elizabeth* (1934, reissued as *Queen Elizabeth I,* 1971). It should be supplemented by other scholarly biographies; among the most useful are J.B. BLACK, *The Reign of Elizabeth, 1558–1603,* 2nd ed. (1959); NEVILLE WILLIAMS, *Elizabeth, Queen of England* (1967; U.S. title, *Elizabeth the First, Queen of England,* 1968), which stresses the formation under Elizabeth of an English national consciousness; and PAUL JOHNSON, *Elizabeth I: A Biography* (U.K. title, *Elizabeth I: A Study in Power and Intellect,* 1974). JASPER RIDLEY, *Elizabeth I* (1987; U.S. title, *Elizabeth I: The Shrewdness of Virtue,* 1988), emphasizes the role of religion in the queen's domestic and foreign policy. Popular biographies of Elizabeth, even when well researched, tend to be highly speculative about Elizabeth's emotions and motivations. Among the more recent biographies are ELIZABETH JENKINS, *Elizabeth the Great* (1958, reissued 1972); LACEY BALDWIN SMITH, *Elizabeth Tudor: Portrait of a Queen* (1975); CAROLLY ERICKSON, *The First Elizabeth* (1983); and ALISON PLOWDEN, *The Young Elizabeth* (1971), and *Elizabeth Regina: The Age of Triumph, 1588–1603* (1980). Selections and extracts of contemporary accounts of Elizabeth may be found in JOSEPH M. LEVINE (ed.), *Elizabeth I* (1969); RICHARD L. GREAVES (ed.), *Elizabeth I, Queen of England* (1974); and LACEY BALDWIN SMITH (ed.), *Elizabeth I* (1980).

Elizabethan government and politics: For the controversy over women's right to rule a nation, see PAULA LOUISE SCALINGI, "The Scepter or the Distaff: The Question of Female Sovereignty, 1515–1607," *Historian,* 41(1):59–75 (1978). The doctrine of the king's two bodies is explained in ERNST H. KANTOROWICZ, *The King's Two Bodies: A Study in Mediaeval Political Theology* (1957, reissued 1987); and applied to the case of Elizabeth in MARIE AXTON, *The Queen's Two Bodies: Drama and the Elizabethan Succession* (1977). ALLISON HEISCH, "Queen Elizabeth I: Parliamentary Rhetoric and the Exercise of Power," *Signs,* 1(1):31–55 (Autumn 1975), analyzes the strategies and effects of Elizabeth's masterful parliamentary speeches.

The structure and practice of Tudor administration is analyzed in PENRY WILLIAMS, *The Tudor Regime* (1979, reissued 1981), which may be supplemented by CHRISTOPHER COLEMAN and DAVID STARKEY (eds.), *Revolution Reassessed: Revisions in the History of Tudor Government and Administration* (1986); and DAVID LOADES, *The Tudor Court* (1986). The operations of Elizabeth's government are treated in detail in WALLACE MACCAFFREY, *The Shaping of the Elizabethan Regime* (1968, reissued 1971), which addresses the early years of her reign, and *Queen Elizabeth and the Making of Policy, 1572–1588* (1981). JOEL HURSTFIELD, *Elizabeth I and the Unity of England* (1960, reissued 1971), deals with Elizabeth's largely successful efforts at creating national unity in the face of profound religious, social, and political changes. For the ways in which Elizabethan politics led to 17th-century revolution, see LAWRENCE STONE, *The Causes of the English Revolution, 1529–1642,* 2nd ed. (1986); and CHRISTOPHER HILL, *Intellectual Origins of the English Revolution* (1965, reprinted 1980).

Aspects of the succession question are addressed by MORTIMER LEVINE, *The Early Elizabethan Succession Question, 1558–1568* (1966); and by JOEL HURSTFIELD, "The Succession Struggle in Late Elizabethan England," in S.T. BINDOFF, JOEL HURSTFIELD, and C.H. WILLIAMS, *Elizabethan Government and Society* (1961), ch. 13, pp. 369–396. Elizabeth's religious policies are studied in WILLIAM P. HAUGAARD, *Elizabeth and the English Reformation: The Struggle for a Stable Settlement of Religion* (1968). The religious affiliations of her councillors are addressed in WINTHROP S. HUDSON, *The Cambridge Connection and the Elizabethan Settlement of 1559* (1980). For foreign policy, see R.B. WERNHAM, *The Making of Elizabethan Foreign Policy, 1558–1603* (1980); and GARRETT MATTINGLY, *Renaissance Diplomacy* (1955, reprinted 1988).

Useful overviews of Elizabethan government are given in ALAN G.R. SMITH, *The Government of Elizabethan England* (1967); S.T. BINDOFF, *Tudor England* (1950, reprinted 1979); and CHRISTOPHER HAIGH (ed.), *The Reign of Elizabeth I* (1984).

The image of Elizabeth: The iconography of the queen's image is examined in ROY STRONG, *Gloriana: The Portraits of Queen Elizabeth I,* rev. ed. (1987), and *The Cult of Elizabeth: Elizabethan Portraiture and Pageantry* (1977, reprinted 1986), which also treats the chivalric revival under Elizabeth. Elizabeth's image in literature is exhaustively treated in ELKIN CALHOUN WILSON, *England's Eliza* (1939, reissued 1966); and in FRANCES YATES, *Astraea: The Imperial Theme in the Sixteenth Century* (1975, reissued 1985). The relations between Elizabeth's image and literary representations of her are investigated in LOUIS ADRIAN MONTROSE, " 'Eliza, Queene of shepheardes,' and the Pastoral of Power," *English Literary Renaissance,* 10(2):153–182 (1980), and " 'Shaping Fantasies': Figurations of Gender and Power in Elizabethan Culture," *Representations,* 1(2):61–94 (Spring 1983). For the staging of the self in this period, see STEPHEN GREENBLATT, *Renaissance Self-Fashioning: From More to Shakespeare* (1980).

Bibliography: A comprehensive bibliography of works relating to Elizabeth and her times is the *Bibliography of British History: Tudor Period 1485–1603,* ed. by CONYERS READ, 2nd ed. (1959, reissued 1978). More recent bibliographies include MORTIMER LEVINE, *Tudor England 1485–1603* (1968); and G.R. ELTON, *Modern Historians on British History, 1485–1945: A Critical Bibliography, 1945–1969* (1970).

(S.J.G.)

Human Emotion

An emotion, as it is commonly known, is a distinct feeling or quality of consciousness, such as joy or sadness, that reflects the personal significance of an emotion-arousing event. In modern times the subject of emotion has become part of the subject matter of several scientific disciplines—biology, psychology, psychiatry, anthropology, and sociology. Emotions are central to the issues of human survival and adaptation. They motivate the development of moral behaviour, which lies at the very root of civilization. Emotions influence empathic and altruistic behaviour, and they play a role in the creative processes of the mind. They affect the basic processes of perception and influence the way humans conceive and interpret the world around them. Evidence suggests that emotions shape many other aspects of human life and human affairs. Clinical psychologists and psychiatrists often describe problems of adjustment and types of psychopathology as "emotional problems," mental conditions that an estimated 1 in 3 Americans, for example, suffers from during his or her lifetime.

The subject of emotion is studied from a wide range of views. Behaviorally oriented neuroscientists study the neurophysiology and neuroanatomy of emotions and the relations between neural processes and the expression and experience of emotion. Social psychologists and cultural anthropologists study similarities and differences among cultures by the way emotions are expressed and conceptualized. Philosophers are interested in the role of emotions in rationality, thought, character development, and values. Novelists, playwrights, and poets are interested in emotions as the motivations and defining features of fictional characters and as vehicles for communicating the meaning and significance of events.

This article will consider the meaning of emotions; the use of emotion concepts in literature and philosophy; the activation, structure, and functions of emotions as conceived by psychologists and neuroscientists; and the causes and consequences of emotions as reflected in individual experience and social relationships.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 433.

The article is divided into the following parts:

Definitions and humanistic background

DEFINITIONS

Emotion has been defined as a particular psychological state of feeling, such as fear, anger, joy, and sorrow. The feeling often includes action tendencies and tends to trigger certain perceptual and cognitive processes. Most experts agree that emotion is a causal factor or influence in thoughts, actions, personalities, and social relationships.

The multilevel concept

The concept of emotion that will be developed here is a multiaspect, or multilevel, one, considering structure and functions at the levels of neurophysiology, emotion expression, and emotion experience (feeling). It should be noted, however, that not all of the numerous definitions that can be found in emotion literature fit into this multilevel concept. The definitions, which reflect differences in the interests and theoretical orientations of the authors, can be reduced to three categories concerned with structure and three concerned with functions. The three structural categories are the three levels, or aspects, that are included in the multilevel concept. The first of these categories of definition focuses on the neurophysiological processes underlying or accompanying emotions, the second on expression, or emotional behaviour, and the third on the subjective experience, or conscious aspect, of emotion.

Of the three categories of definition related to functions, the first defines emotions in terms of their adaptive or disruptive influences. The second category defines emotion in terms of motivation and considers it as part of the same class of phenomena that contains physiological drives, such as pain, thirst, and the need for elimination. The third category concerned with functions consists of definitions that attempt to distinguish between emotion and other psychological processes.

A multilevel definition of emotion essentially subsumes definitions that focus on one of the three structural categories of neural processes, expressive behaviour, and subjective experience, and elaborations and extensions of such a definition would consider concerns of the three categories related to functions. In summary, the foregoing consideration of definitions of emotion suggests that a multilevel concept comes closest to a consensus viewpoint among emotion theorists and provides a way of resolving the complex issue of definition. Thus, a specific emotion is a particular set of neural processes that gives rise to a particular configuration of expressive behaviours and a particular feeling state or quality of consciousness that has motivational and adaptive functions. Under some circumstances extremely intense emotion may become disruptive.

HUMANISTIC BACKGROUND

Orators, literary artists, and philosophers have recognized emotions as part of human nature since recorded history. Homer's *Iliad* contains vivid descriptions of the emotions of the characters; the goddess Athena frequently goes among Agamemnon's troops playing upon their emotions, attempting to allay their fears and bolster their courage for battle. Ancient philosophers discussed the emotions at length, and from these discussions it appears that the basic meanings of emotion concepts are timeless. For example,

Aristotle's views

in the *Rhetoric,* Aristotle described the significance, causes, and consequences of the experiences of anger, fear, and shame in much the same way as contemporary writers. He observed that anger is caused by undeserved slight, fear by the perception of danger, and shame by deeds that bring disgrace or dishonour. His understanding of the relations among emotions also has a modern ring. In contrasting the young and the old, he said of the young,

> And they are more courageous, for they are full of passion and hope, and the former of these prevents them fearing, while the latter inspires them with confidence, for no one fears when angry, and hope of some advantage inspires confidence.

Literature. The use of emotion words in literary works serves several purposes. They help define the motivations and personalities of the characters in a play or novel, and they help the reader to understand and identify with characters and to experience vicariously their emotions.

Shakespeare, for example, was a master at expressing emotion through his characters and eliciting emotions from the audience. His work also contains quite accurate descriptions of emotional expressions. An example in *Henry V* is the king's effort to ready his soldiers for battle:

> Then imitate the action of the tiger;
> Stiffen the sinews, summon up the blood,
> Disguise fair nature with hard-favour'd rage;
> Then lend the eye a terrible aspect;
> Let it pry through the portage of the head
> Like the brass cannon; let the brow o'erwhelm it
> As fearfully as doth a galled rock
> O'erhang and jutty his confounded base,
> Swill'd with the wild and wasteful ocean.
> Now set the teeth and stretch the nostril wide,
> Hold hard the breath and bend up every spirit
> To his full height.
>
> (Act III, scene 1)

James Joyce's use of emotion

In modern times James Joyce used emotion words and words with emotional connotation to powerful effect. In *A Portrait of the Artist as a Young Man,* much of Stephen Dedalus' mood and character are revealed in a few lines describing a time when he was drinking with his cronies and trying to overcome his sense of alienation from his father:

> His mind seemed older than theirs: it shone coldly on their strifes and happiness and regrets like a moon upon a younger earth. . . . Nothing stirred within his soul but a cold and cruel and loveless lust. His childhood was dead or lost and with it his soul capable of simple joys, and he was drifting amid life like the barren shell of the moon.

According to the literary critic Rosemarie Battaglia, the emotion-arousing words cold, cruel, loveless, dead, lost, and barren resonate with a sense of Stephen's withdrawal from his social world.

Other modern writers have made frank use of psychological concepts of emotion and emotion-related processes, particularly those introduced by Sigmund Freud. Thus, for example, the author's characters may be motivated by unconscious processes, feelings they cannot label and articulate because the fundamental underlying ideation associated with the feelings has been repressed.

Philosophy. Using Aristotle's system of causal explanation, the 16th-century British philosopher John Rainolds defined emotion as follows: the efficient cause of emotions is God, who implanted them; the material cause is good and evil human things; the formal cause is a commotion of the soul, impelled by the sight of things; and the final cause is seeking good and fleeing evil. The American philosopher L.D. Green's commentary on Rainolds' thesis indicates that Rainolds was not faithful to Aristotle's own discussions of emotion.

One thing that Aristotle did advocate was moderation of emotions, allowing them to have an effect only at the right time and in the right manner. Rainolds noted that the Aristotelian thinker Cicero saw emotions as beneficial—fear making humans careful, compassion and sadness leading to mercy, and anger whetting courage. These thoughts about emotion are similar to those of some modern theorists.

For Rainolds, the emotions are the active, energizing aspects of human nature. Although the intellect exercises control over emotions, intellect can have no impact without emotion. Rainolds was specifically concerned with the effects of emotion on rhetoric, but he saw rhetoric as a principal means of influencing human behaviour and affairs. He believed that

> the passions [emotions] must be excited, not for the harm they do but for the good, not so they twist the straight but that they straighten the crooked; so they ward off vice, iniquity, and disgrace; so that they defend virtue, justice, and probity.

Spinoza's idea of emotion and bodily change

Benedict de Spinoza in the 17th century described emotions in much the same way as Rainolds did, but he discussed them in relation to action rather than to language. He saw emotions as bodily changes that result in the amplification or attenuation of action and as processes that can facilitate or impede action. For Spinoza, emotion also included the ideas, or mental representations, of the bodily changes in emotion.

Blaise Pascal and David Hume reversed Rainolds' position by assuming the primacy of emotion in human behaviour. Hume said that reason is the slave of the passions (emotions), and Pascal observed in *Pensées* that "the heart has reasons that reason does not know." Although Hume believed that passions (emotions) rule reason or intellect, he thought the dominant passion should be moral sentiment. Some contemporary psychologists trace morality to empathy and empathy to discrete emotions including sadness, sorrow, compassion, and guilt.

Since Rainolds lectured on emotions at Oxford, philosophers have considered many questions related to emotions: Are they active or passive? Can they be explained by neurophysiological processes and reduced to material phenomena? Are they rational or nonrational? Are they voluntary or involuntary? Characterizing or categorizing emotions according to these dichotomies has resulted in yet other classifications or distinctions.

Ultimately, emotion concepts resist definition by way of dichotomous distinctions. Emotions are generally active and tend to generate action and cognition, but extreme fear may cause behavioral freezing and mental rigidity. Emotion can be explained on one level in terms of neurochemical processes and on another level in terms of phenomenology. Emotions are rational in the sense that they serve adaptive functions and make sense in terms of the individual's perception of the situation. They are nonrational in the sense that they can exist in the brain at the neurochemical level and in consciousness as unlabeled feelings that may be independent of cognitive-rational processes. Emotions are voluntary in that their expression in older children and adults is subject to considerable modification and control via cognition and action, and willful regulation of expression may result in regulation of emotion experience. Emotions are involuntary in that an effective stimulus elicits them automatically, without deliberation and conscious choice. Nowhere is this more evident than in infants and young children, who have little capacity to modulate or inhibit emotion by means of cognitive processes.

Rational and nonrational emotions

One contemporary American philosopher, Amélie O. Rorty, espouses a three-part causal history for emotions, which includes (1) the formative events in a person's past, including the development of habits of thought, (2) sociocultural factors, and (3) genetically determined sensitivities and patterns of response. These are essentially the same factors that are recognized by psychologists, who frequently reduce the list to two: (1) experience as mediated by culture and learning and (2) genetic determinants that unfold with ontogenetic development. The first of these two causal factors indicates that individual differences in interpretations of an event or situation lead to different emotions in different persons.

Some philosophers are concerned with the question of the rationality of emotion as judged on the basis of causes and consequences. One resolution is in terms of appropriateness: an emotion is appropriate if the reasons for it are adequate, regardless of the reasons against it. There may be a sense, however, in which emotions are intrinsically nonrational because they can come into a person's consciousness without that person having considered all of the relevant reasons for them. In the final analysis, caution should be used in judging the rationality of emotions.

James Hillman's conception

Another contemporary philosopher, James Hillman, has been notably effective in using classical philosophical principles to explain emotions. He has delineated 12 ways that emotion has been conceptualized in philosophy and psychology. These include conceptions of emotion as a distinct entity or trait, an accompaniment of instinct, energy for thought and action, a neurophysiological mechanism and process, mental representation, signal, conflict, disorder, and creative organization. This philosopher found each of these conceptions incomplete or incorrect and returned to Aristotle's system of four causes in an effort to integrate the information from each of the foregoing approaches to defining and studying emotions.

For Hillman, the efficient cause of emotion, described psychologically, consists of conscious or unconscious mental representations (perceptions, images, or thoughts) and conflicts between physiological or psychological systems or between a person and the environment. The efficient cause described physiologically includes genetic endowment and the neurochemical and hormonal processes involved in emotion activation. Hillman stated that the material cause of emotion is energy. He argued that matter, the ultimate source of energy, is relative and that emotion, as the psychological aspect of general energy, is going on all the time and is a two-way bridge uniting subject and object.

In considering the formal cause, one may see emotion as a pattern of neurophysiological and expressive behaviours and subject–object relations. Hillman concluded that, in a formal sense, emotion is a total pattern of the soul:

> Emotion is the soul as a complex whole, involving constitution, gross physiology, facial expression in its social context as well as actions aimed at the environment.

The final cause, or purpose, of emotion, according to Hillman, can be thought of in terms of what it achieves: survival (energy release, homeostatic regulation, and action on the stimulus and environment), signification (qualification of experience, expression, communication, and values), and improvement (emergence of energy into consciousness, facilitation of creative activity, and strengthening of the organization of self and behaviour). Hillman integrated these various descriptions of final cause in the concept of change. Emotion occurs in order to actualize change; "emotion itself is change."

HOW PSYCHOLOGY CONCEIVES EMOTIONS

In 1872, emotion studies received a boost in scientific status when Charles Darwin published his seminal treatise *The Expressions of the Emotions in Man and Animals.* Twelve years later, the American philosopher and psychologist William James, one of the pioneers of psychology in the United States, published what was to become a famous and controversial theory of emotions. In it James proposed that an arousing stimulus (such as a poignant memory or a physical threat) triggers internal physiological processes as well as external expressive and motor actions and that the feeling of these physiological and behavioral processes constitutes the emotion. Thus, people are happy because they smile, sad because they cry, angry because they frown, and afraid because they run from danger.

A few years later the Danish physician Carl Lange published a more constricted theory, maintaining that emotion is a function of the perception of changes in the visceral organs innervated by the autonomic nervous system. Although there were distinctively individual components in the theories of James and Lange, the theories became linked in the minds of psychologists and the combination became known as the James–Lange theory.

The James–Lange theory

The James–Lange theory was seriously challenged by the American physiologist Walter B. Cannon, who showed that, among other things, animals whose viscera were separated from the central nervous system still displayed emotion expression. Cannon contended that bodily changes were similar for most kinds of emotions, whereas the James–Lange theory implied a different bodily pattern of response for different emotions. The James–Lange theory has remained a more or less permanent fixture in behavioral science nevertheless, and most psychology textbooks summarize the theory and Cannon's criticisms of it. Some theories of emotion are classified as neo-Jamesian, and most theories can be identified or classified on the basis of their similarities and differences with the landmark James–Lange theory.

Psychological theories of emotion can be grouped into two broad categories—biosocial and constructivist. Although this system of categorization is an oversimplification, it provides a way for the student of emotion to get a perspective on a particular theory. A contemporary textbook, for example, describes 20 psychological theories of emotion, and there are many others that it does not consider.

Many of the differences between the two categories of emotion theory stem from different assumptions regarding the relative importance of genetics and life experiences. Biosocial theories assume that emotions are rooted in biological makeup and that genes are significant determinants of the threshold and characteristic intensity level of each basic emotion. In this view, emotional life is a function of the interaction of genetic tendencies and the evaluative systems, beliefs, and roles acquired through experience. Constructivist theories assume that genetic factors are inconsequential and that emotions are cognitively constructed and derived from experience, especially from social learning. The constructivists' crucible for emotions is formed by the interactions of the person with the environment, especially the social environment. Thus, according to the constructivists, emotions are a function of appraisals, or evaluations, of the world of culture, and of what is learned. (For examples of both types of theory and some of the research generated by each, see below *Contemporary approaches to emotion.*)

Biosocial and constructivist theories

The importance of emotions

The use of emotion concepts is common in literature and philosophy, as was discussed above, and there is widespread agreement among scientists that emotions are important in individual development, physical and mental health, and human relations. Experts in different disciplines emphasize different reasons for the importance of emotions.

EVOLUTIONARY-BIOLOGICAL PERSPECTIVES

Darwin and emotion expressions

Darwin included emotions, in particular emotion expressions, in his studies of evolution. He considered continuity or similarity of expression in animals and human beings as further evidence of human evolution from lower forms. His finding that certain emotion expressions are innate and universal was seen as evidence of the "unity of the several races." Thus, the expressions, or the language of the emotions, provide a means of communication among all human beings, regardless of culture or ethnic origin.

In his work *The Expression of the Emotions in Man and Animals,* Darwin made an explicit value judgment regarding the significance of emotion expressions:

> The movements of expression in the face and body, whatever their origin may have been, are in themselves of much importance for our welfare. They serve as the first means of communication between the mother and infant; she smiles approval, and thus encourages her child on the right path, or frowns disapproval. We readily perceive sympathy in others by their expression; our sufferings are thus mitigated and our pleasures increased; and mutual good feeling is thus strengthened. The movements of expression give vividness and energy to our spoken words. They reveal the thoughts and intentions of others more truly than do words, which may be falsified.

From his studies of emotion expressions, Darwin concluded that some emotion expressions were due to the "constitution of the nervous system," or our biological endowment. The implication is that these expressive movements are part of human nature and have played a role in survival and adaptation. Darwin thought other expressions were derived from actions that originally served biologically adaptive functions (*e.g.,* preparation for biting became the bared teeth of the anger expression). Although he noted that expressive movements may no longer serve biological functions, he made it quite clear that they serve critical social and communicative functions.

PSYCHOLOGICAL VIEWS

Significance of emotions

From the very beginning of scientific psychology, there were voices that spoke of the significance of emotions for human life. James believed that "individuality is founded in feeling" and that only through feeling is it possible "directly to perceive how events happen, and how work is actually done." The Swiss psychiatrist Carl Gustav Jung recognized emotion as the primal force in life:

> But on the other hand, emotion is the moment when steel meets flint and a spark is struck forth, for emotion is the chief source of consciousness. There is no change from darkness to light or from inertia to movement without emotion.

Psychologists did not rally to the Darwinian thesis on the evolutionary-adaptive functions of emotions in significant numbers until the 1960s. Several influential volumes following this theme were published in the 1960s and '70s. For example, the American psychologist Robert Plutchik echoed Darwinian principles in several of the postulates of his theory: emotions are present at all levels of animal life, and they serve an adaptive role in relation to survival issues posed by the environment.

The American psychologist Silvin Tomkins believed that the emotions constitute the primary motivational system for human beings. He held that even physiological drives such as hunger and sex obtain their power from emotions and that the energizing effects of emotion are necessary to sustain drive-related actions. In this way, he argued that emotions are essential to survival and adaptation.

Other theorists and researchers that follow the Darwinian principles of the survival value and adaptive value of emotions have emphasized their role in human development and in the development of social bonds, particularly mother–infant or parent–child attachment. These researchers have shown that even the very young infant has a repertoire of emotion expressions translatable into messages calling for nourishment and affection, both essential ingredients of healthy development. The distress expression is the infant's all-out cry for help, the sadness expression an appeal for empathy, and the smile an invitation to stimulating face-to-face interactions. (For discussion of empirical evidence of the importance of emotions in child development, social relations, cognitive processes, and mental health, see below *The functions of emotion.*)

Infant emotions

Contemporary approaches to emotion

Contemporary psychologists are concerned with the activation, or causes, of emotion, its structure, or components, and its functions or consequences. Each of these aspects can be considered from both a biosocial and a constructivist view. On the whole, biosocial theories have been relatively more concerned with the neurophysiological aspects of emotions and their roles as motivators and organizers of cognition and action. Constructivists have been relatively more concerned with explaining the causes of emotion at the experiential level and cognition-emotion relations in terms of cognitive–linguistic processes.

STRUCTURES AND PROCESSES OF EMOTION ACTIVATION

The question of precisely how an emotion is triggered has been one of the most captivating and controversial topics in the field. To address the question properly, one must break it down into more precise parts. Emotion activation can be divided into three parts: neural processes, bodily (physiological) changes, and mental (cognitive) activity.

While it is easy for people to think of things that make them happy or sad, it is not yet possible to explain precisely how the feelings of joy and sadness occur. Neuroscience has produced far more information about the processes leading to the physiological responses and expressive behaviour of emotion than about those that generate the conscious experience of emotion.

Neural processes. An emotion can be activated by causes and processes within the individual or by a combination of internal and external causes and processes. For example, within the individual, an infection can cause pain, and pain can activate anger.

The findings of neuroscience indicate that stimuli are evaluated for emotional significance when information from primary receptors (in the visual, tactual, auditory, or other sensory systems) travels along certain neural pathways to the limbic forebrain. Scientific data developed by Joseph E. LeDoux show that auditory fear conditioning involves the transmission of sound signals through the auditory pathway to the thalamus (which relays information) in the lower forebrain and thence to the dorsal amygdala (which evaluates information).

The brain's involvement

Evidence from neuroscience suggests that emotion activated by way of the thalamo-amygdala (subcortical) pathway results from rapid, minimal, automatic, evaluative processing. Emotion activated in this way need not involve the neocortex. Emotion activated by discrimination of stimulus features, thoughts, or memories requires that the information be relayed from the thalamus to the neocortex. Such a circuit is thought to be the neural basis for cognitive appraisal and evaluation of events.

This two-circuit model of the neural pathways in emotion activation has several important theoretical implications. The neurological evidence indicating that emotion can be activated via the thalamo-amygdala pathway is consistent with the behavioral evidence that very young infants respond emotionally to pain and that adults can develop preferences or make affective judgments in responding to objects before they demonstrate recognition memory for them. This suggests that in some instances humans may experience emotion before they reason why.

It might be expected that in early human development most emotion expressions derive from automatic, subcortical processing, with minimal cortical involvement. As cognitive capacities increase with maturation and learning, the neocortex and the cortico-amygdala pathway become more and more involved. By the time children acquire language and the capacity for long-term memory, they may process events in either or both pathways, with the subcortical pathway specializing in events requiring rapid response and the cortico-amygdala pathway providing evaluative information necessary for cognitive judgment and more complex coping strategies.

Physiological processes. Many theorists agree that feedback from physiological activity contributes to emotion activation. There is disagreement over the kind of feedback that is important. Some think that it is a visceral feedback—coming from the activity of the smooth-muscle organs such as the heart and stomach, which are innervated by the autonomic nervous system. Others believe that it is feedback from the voluntary, striated muscles, especially of the face, which are innervated by the somatic nervous system.

Cognitive processes. Constructivist theorists and researchers have been concerned with the causes of emotion at the cognitive-experiential level and with the relations between cognitive processes and emotion. This research has focused on two topics: the relations between appraisals, or evaluations, and emotions and the relations between causal attributions and emotions.

Magda B. Arnold was the first contemporary psychologist to propose that all emotions are a function of one's cognitive appraisal of the stimulus or situation. She maintained that before a stimulus can elicit emotion it has to be appraised as good or bad by the perceiver. She described the appraisal that arouses emotion as concrete, immediate, undeliberate, and not the result of reflection. Her position was adopted and elaborated by others, some of whom assumed that cognitive activity, whether in the form of primitive evaluative perception or symbolic processes, is a necessary precondition of emotion. Biosocial and constructivist theorists agree that cognition is an important determinant of emotion and that emotion–cognition relations merit continued research.

Emotion appraisal

Research by the American psychologists Phoebe C. Ellsworth and Craig A. Smith on the relations between appraisals and specific emotions show that people tend to appraise situations in terms of elements such as pleasantness, anticipated effort, certainty, responsibility, control, legitimacy, and perceived obstacle. Researchers have found that each discrete emotion tends to be associated with a distinctive combination of appraisals. For example, a perceived obstacle (barrier to a goal) that is due to

someone else's responsibility is associated with anger, a perceived obstacle that is the person's own responsibility is associated with guilt, and a perceived obstacle characterized by uncertainty is associated with fear. This study was based on subjects' retrospective accounts of emotion-eliciting situations, and therefore the data cannot confirm the view that appraisal causes emotion. However, the assumption that emotion and appraisal are causally related seems reasonable.

Another approach to explaining the causes of emotions is that of attribution theory. The central idea of this theory, according to the American psychologist Bernard Weiner, is that the perceptions of the causes of events can be characterized in three principal ways which affect many emotional experiences. The perceived causes of events (*e.g.,* success and failure) are characterized by their locus (internal or external to the person), stability (a trait of the person or a temporary condition), and controllability (under the person's control or not).

Attribution theory

Research has shown that different patterns of causal attribution are associated with different emotions, including anger, guilt, shame, and the more complex phenomena of pity, pride, gratitude, and hopelessness. Pity is attributed to the perception of uncontrollable and stable causes—people feel pity for a person who has an affliction due to a genetic defect or accident. Anger is attributed to external and controllable events—people feel anger when an affront or injury is caused by someone's lack of concern or thoughtlessness. Guilt is attributed to the perception of internal and controllable causes—people feel guilt for wrongdoing they could have avoided. Children aged five to 12 understand the emotional consequences of revealing the causes of their actions; they know that their teachers might be angry at their failure if they have not tried hard enough and that teachers might feel pity for students who lack the ability to learn efficiently and perform well.

Coping with emotions

Psychologists researching cognitive activation have studied the relations between the ways people cope with stressful encounters and the emotions they experience after their efforts to resolve the problems. In one study emotions were assessed by asking subjects to indicate the extent to which they experienced emotions on four scales: worried/fearful, disgusted/angry, confident, and pleased/happy. Coping was assessed by subjective ratings on eight scales: confrontive coping ("stood my ground and fought"), distancing ("didn't let it get to me"), self-control ("tried to keep my feelings to myself"), seeking social support ("talked to someone"), accepting responsibility ("criticized myself"), escape-avoidance ("wished the situation would go away"), planful problem solving ("changed or grew as a person"), and positive reappraisal. Four of these ways of coping were associated with the quality of emotion that followed the effort to cope. Planful problem solving and positive reappraisal tended to increase happiness and confidence and to decrease disgust and anger. Obversely, the subjects reported that confrontation and distancing techniques increased their disgust and anger and decreased their happiness and confidence. Because these data were retrospective, there can be no firm conclusion that a particular way of coping causes a particular emotion experience. Nevertheless, the observed relations among ways of coping and subsequent emotion experiences are reasonable and in line with theoretical expectations.

The controversy as to whether some cognitive process is a necessary antecedent of emotion may hinge on the definition of terms, particularly the definition of cognition. If cognition is defined so broadly that it includes all levels or types of information processing, then cognition may confidently be said to precede emotion activation. If those mental processes that do not involve mental representation based on learning or experience are excluded from the concept of cognition, then cognition so defined does not necessarily precede the three-week-old infant's smile to the high-pitched human voice, the two-month-old's anger expression to pain, or the formation of the affective preferences (likes or dislikes) in adults.

Multimodal theory. Evidence suggests that a satisfactory model of emotion activation must be multimodal. Emotions can, as indicated above, be activated by such pre-cognitive processes as physiological states, motor mimicry (imitation of another's movements), and sensory processes and by numerous cognitive processes, including comparison, matching, appraisal, categorization, imagery, memory, attribution, and anticipation. Further, all emotion activation processes are influenced by a variety of internal and external factors.

THE STRUCTURE OF EMOTIONS

In the discussion of the structure of emotions it is not always possible to ignore the function of emotions, which is discussed in the following section. The separation, however, is conducive to sorting out the complex field of emotions.

Both biosocial and constructivist theories of emotions acknowledge that an emotion is a complex phenomenon. They generally agree that an emotion includes physiological functions, expressive behaviour, and subjective experience and that each of these components is based on activity in the brain and nervous system. As noted above, some theorists, particularly those of the constructivist persuasion, hold that an emotion also involves cognition, an appraisal or cognitive-evaluative process that triggers the emotion and determines or contributes to the subjective experience of the emotion.

Biosocial and constructivist views

The physiological component. The physiological component of emotion has been a lively topic of research since Cannon challenged the James–Lange theory by showing that feedback from the viscera has little effect on emotional expression in animals. Cannon's studies and criticisms were regarded by many as too narrow, failing to, among other things, consider the possible role of feedback from striated muscle systems of the face and body.

Role of the nervous system. Since the popularization of the James–Lange theory of emotion, the physiological component of emotion has been traditionally identified as activity in the autonomic nervous system and the visceral organs (*e.g.,* the heart and lungs) that it innervates. However, some contemporary theorists hold that the neural basis of emotions resides in the central nervous system and that the autonomic nervous system is recruited by emotion to fulfill certain functions related to sustaining and regulating emotion experience and emotion-related behaviour. Several findings from neuroscience support this idea. Neuroanatomical studies have shown that the central nervous system structures involved in emotion activation can exert direct influences on the autonomic nervous system. For example, efferents from the amygdala to the hypothalamus may influence activity in the autonomic nervous system that is involved in defensive reactions. Further, there are connections between pathways innervating facial expression and the autonomic nervous system. Studies have shown that patterns of activity in this system vary with the type of emotion being expressed.

Findings of neuroscience

Roles of the brain hemispheres. There is some evidence that the two hemispheres of the brain are related differently to emotion processes. Early evidence suggested that the right (or dominant) hemisphere may be more adept than the left at discriminating among emotional expressions. Later research using electroencephalography elaborated this initial conclusion, suggesting that the right hemisphere may be more involved in processing negative emotions and the left hemisphere more involved in processing positive emotions.

The expressive component. The expressive component of emotion includes facial, vocal, postural, and gestural activity. Expressive behaviour is mediated by phylogenetically old structures of the brain, which is consistent with the notion that they served survival functions in the course of evolution.

Involvement of brain structures. Emotion expressions involve limbic forebrain structures and aspects of the peripheral nervous system. The facial and trigeminal nerves and receptors in facial muscles and skin are required in expressing emotion and in facilitating sensory feedback from expressive movements.

Early studies of the neural basis of emotion expression showed that aggressive behaviour can be elicited from a cat after its neocortex has been removed and suggested

that the hypothalamus is a critical subcortical structure mediating aggression. Later research indicated that, rather than the hypothalamus, the central gray region of the midbrain and the substantia nigra may be the key structures mediating aggressive behaviour in animals.

Importance of facial expression

Neural pathways of facial expression. Of the various types of expressive behaviour, facial expression has received the most attention. In human beings and in many nonhuman primates, patterns of facial movements constitute the chief means of displaying emotion-specific signals. Whereas research has provided much information on the neural basis of emotional behaviours (*e.g.*, aggression) in animals, little is known about the brain structures that control facial expression.

The peripheral pathways of facial emotion expression consist of the seventh and fifth cranial nerves. The seventh, or facial, nerve is the efferent (outward) pathway; it conveys motor messages from the brain to facial muscles. The fifth, or trigeminal, nerve is the afferent (inward) pathway that provides sensory data from movements of facial muscles and skin. According to some theorists, it is the trigeminal nerve that transmits the facial feedback which contributes to the activation and regulation of emotion experience. The impulses for this sensory feedback originate when movement stimulates the mechanoreceptors in facial skin. The skin is richly supplied with such receptors, and the many branches of the trigeminal nerve detect and convey the sensory impulses to the brain.

The innateness and universality of emotion expressions. More than a century ago Darwin's observations and correspondence with friends living in different parts of the world led him to conclude that certain emotion expressions are innate and universal, part of the basic structure of emotions. Contemporary cross-cultural and developmental research has given strong support to Darwin's conclusion, showing that people in literate and preliterate cultures have a common understanding of the expressions of joy, surprise, sadness, anger, disgust, contempt, and fear. Other studies have suggested that the expressions of interest and shyness and the feelings of shame and guilt may also be innate and universal.

The experiential component. There is general agreement that various stimuli and neural processes leading to an emotion result not only in physiological reactions and expressive behaviour but also in subjective experience. Some biosocial theorists restrict the definition of an emotion experience to a feeling state and argue that it can be activated independently of cognition. Constructivist theorists view the experiential component of emotion as having a cognitive aspect. The issue regarding the relation between emotion feeling states and cognition remains unresolved, but it is widely agreed that emotion feeling states and cognitive processes are typically highly interactive.

Emotion experiences, the actual feelings of joy, sadness, anger, shame, fear, and the like, do not lend themselves to objective measurement. All research on emotion experience ultimately depends on self-reports, which are imprecise. There are few instances where feelings and words are perfectly matched. Yet, most students of emotions, whether philosopher or neuroscientist, ultimately want to explain emotion experience.

Neural basis of emotion experience

The physiological structure of emotion experience. Little is known about the neural basis of emotion experience. Critical reviews have shown that there is little evidence to support the position that activity in the autonomic nervous system provides the physiological basis for emotion experience. However, there is some evidence to support the hypothesis that sensory feedback from facial expression contributes to emotion experience.

Cognitive models of emotion experience have influenced conceptions of the underlying neural processes. Explanations of emotions in terms of appraisal and attributional processes led some researchers to suggest that conscious experiences of emotions derive from the cognitive processes that underlie language. This led to the hypothesis that emotion experiences involve interactions between limbic forebrain areas and the areas of the neocortex that mediate language and language-based cognitive systems. However, this view does not take into account the possibility that emotions occur in preverbal infants and may be mediated in adults by unconscious or nonlinguistic mental processes, such as imagery.

Action tendencies in emotion experiences. Both constructivist and biosocial theorists have emphasized that emotions include action tendencies. The experience, or feeling, of a given emotion generates a tendency to act in a certain way. For example, in anger the tendency is to attack and in fear to flee. Whether a person actually attacks in anger or flees in fear depends on the individual's methods of emotion regulation and the circumstances.

THE FUNCTIONS OF EMOTIONS

In academic discussions of the functions of emotions the focus is usually on the phenomenological, or experiential, aspect of emotions. For purposes of this discussion, however, the functions of emotions are examined in terms of the three structural components—physiological, expressive, and experiential.

Structural components of functions

Physiological functions. The functions of physiological activity that is mediated by the autonomic nervous system and that accompanies states of emotion can be considered as part of the individual's effort to adapt and cope, but, of course, physiological as well as cognitive reactions in extreme emotion usually require regulation (expressed through cognitive processes and expressive behaviour) in order for coping activities to be effective. For example, adaptation to situations that elicit a less extreme emotion such as interest require a quite different physiological and behavioral activity than do situations that elicit intense anger or fear. The heart-rate deceleration and quieting of internal organs that occur in interest facilitate the intake and processing of information, whereas heart-rate acceleration in intense anger and fear prepares the individual to cope by more active means, whether through shouting, physical actions, or various combinations of the two.

Functions of emotion expressions. Emotion expressions have three major functions: they contribute to the activation and regulation of emotion experiences; they communicate something about internal states and intentions to others; and they activate emotions in others, a process that can help account for empathy and altruistic behaviour.

Feedback of facial expression

Role of expressions in emotion experiences. In *The Expression of the Emotions in Man and Animals* Darwin clearly revealed his belief that even voluntary emotion expression evoked emotion feeling. He wrote: "Even the simulation [expression] of an emotion tends to arouse it in our minds." Thus, Darwin's idea suggested that facial feedback (sensations created by the movements of expressive behaviour) activate, or contribute to the activation of, emotion feelings. A number of experiments have provided substantial evidence that intentional management of facial expression contributes to the regulation (and perhaps activation) of emotion experiences. Most evidence is related not to specific emotion feelings but to the broad classes of positive and negative states of emotion. There is, therefore, some scientific support for the old advice to "smile when you feel blue" and "whistle a happy tune when you're afraid."

Darwin was even more persuasive when speaking specifically of the regulation of emotion experience by self-initiated expressive behaviour. He wrote:

> The free expression by outward signs of an emotion intensifies it. On the other hand, the repression, as far as this is possible, of all outward signs softens our emotions.

Experiments by more contemporary researchers on motivated, self-initiated expressive behaviours have shown that, if people can control their facial expression during moments of pain, there will be less arousal of the autonomic nervous system and a diminution of the pain experience.

Role of expressions in communicating internal states. The social communication function of emotion expressions is most evident in infancy. Long before infants have command of language or are capable of reasoning, they can send a wide variety of messages through their facial expressions. Virtually all the muscles necessary for facial expression of basic emotions are present before birth. Through the use of an objective, anatomically based system for coding the separate facial muscle movements, it

has been found that the ability to smile and to facially express pain, interest, and disgust are present at birth; the social smile can be expressed by three or four weeks; sadness and anger by about two months; and fear by six or seven months. Informal observations suggest that expressions indicative of shyness appear by about four months and expressions of guilt by about two years.

The expressive behaviours are infants' primary means of signaling their internal states and of becoming engaged in the family and larger human community. Emotion expressions help form the foundation for social relationships and social development. They also provide stimulation that appears to be necessary for physical and mental health.

Mother–infant relations through emotion expression

Role of expressions in motivating response. One- and three-day-old infants cry in response to other infants' cries but not to a computer-generated sound that simulates crying. Infants as young as two or three months of age respond differently to different expressions by the mother. The information an infant obtains from the mother's facial expressions mediates or regulates a variety of infant behaviours. For example, most infants cross a modified "visual cliff" (an apparatus that was originally used in depth perception study, consisting of a glass floor that gives the illusion of a drop-off) if their mother stands on the opposite side and smiles, but none cross if she expresses fear.

Facial expressions, particularly of sadness, may facilitate empathy and altruistic behaviour. Darwin thought facial expressions evoked empathy and concluded that expression-induced empathy was inborn. Research has shown that, when mothers display sadness expressions, their infants also demonstrate more sadness expressions and decrease their exploratory play. Infants under two years of age respond to their mother's real or simulated expressions of sadness or distress by making efforts to show sympathy and provide help.

Functions of emotion experiences. Psychologists who adopt a strong behaviourist position deny that emotion experiences are matters for scientific inquiry. In contrast, some biosocial theories hold that emotion feelings must be studied because they are the primary factors in organizing and motivating human behaviour. According to these theories, most of the functions attributed to emotion expressions, such as empathy and altruism, are dependent on the organizing and motivating properties of underlying emotion feelings. Emotion experiences have several other functions.

Commonality of emotion experience

Research has shown that people in widely different literate and preliterate cultures not only recognize basic emotion expressions but also characterize and label them with semantically equivalent terms. It seems reasonable to assume that the common feeling state of a given emotion generates the cues for the cognitive processes that result in universal emotion concepts. Of course, if researchers include contextual factors, such as societal taboos, in their description of an emotion experience, they then find differences across cultures. In any case, although the feeling of a given emotion, say fear, may be constant, people within and across cultures learn to be afraid of quite different things and to cope with fear in different ways.

Experiential influence on cognitive processes. Several lines of research have shown that induced emotion affects perception, learning, and memory. In one study, conducted by Carroll E. Izard and his students, subjects were made happy or angry and then shown happy and angry faces and friendly and hostile interpersonal scenes in a stereoscope. Happy subjects perceived more happy faces and friendly interpersonal scenes, and angry subjects perceived more angry faces and hostile interpersonal scenes. In this case, emotion apparently altered the basic perceptual process. In another study subjects were made happy or sad and then given happy and sad information about fictional persons and later asked to give their impressions and make judgments about the fictional characters. Overall, happy subjects reported more favourable impressions and positive judgments than did sad subjects. These studies provide evidence for the common wisdom that happy people are more likely to see the world through rose-coloured glasses.

Experiential facilitation of empathy and altruism. An extensive series of studies indicated that positive emotion feelings enhance empathy and altruism. It was shown by the American psychologist Alice M. Isen that relatively small favours or bits of good luck (like finding money in a coin telephone or getting an unexpected gift) induced positive emotion in people and that such emotion regularly increased the subjects' inclination to sympathize or provide help.

Effects of positive emotion

Experiential relation to increased creativity. Several studies have demonstrated that positive emotion facilitates creative problem solving. One of these studies showed that positive emotion enabled subjects to name more uses for common objects. Another showed that positive emotion enhanced creative problem solving by enabling subjects to see relations among objects that would otherwise go unnoticed. A number of studies have demonstrated the beneficial effects of positive emotion on thinking, memory, and action in preschool and older children.

Explanation of the functions of emotion experiences. There are two kinds of factors that contribute to the enhancing effects of positive emotion on perception, learning, creative problem solving, and social behaviour. Two factors, emphasized by cognitive-social theorists, are related to cognitive processes. First, positive emotion cues positive material in memory, and, second, positive material in memory is more extensive than neutral and negative material. The second set of factors, emphasized by biosocial theorists, are related to the intrinsic motivational and organizational influences of emotion and to the particular characteristics of the subjective experience of positive emotion. For example, these theorists maintain that the experience of joy is characterized by heightened self-esteem and self-confidence. These qualities of consciousness increase the receptibility to information and the flexibility of mental processes. Biosocial theorists consider that the positive emotion induced by experimental manipulations and experimental tasks includes the emotion of interest, which is characterized by curiosity and the desire to explore and learn. The concepts emphasized by biosocial and cognitive-social theories may be seen as complementary.

EMOTIONS AND ADAPTATION

The results of many of the experiments discussed above indicate that emotions have motivational and adaptive properties. Perhaps the most convincing demonstrations of this come from studies showing that emotions influence perception, learning, and memory and empathic, altruistic, and creative actions.

Some theorists have viewed emotions more negatively, seeing them as disorganizing and disrupting influences. Researchers in this tradition have also viewed emotions as transient, episodic states. These ideas were fueled by a research emphasis on "emergency emotions," such as rage and panic. These researchers might agree that, although such emotions may serve an adaptive function under certain circumstances, in many situations they can lead to behaviours that prove to be maladaptive and even fatal. As was indicated above, however, emotion expressions can serve critical functions in mother–infant communication and attachment, and emotion experiences, or feeling states, facilitate learning and empathic, altruistic, and creative behaviour.

Biological adaptations

Although psychologists generally favour viewing emotions as having motivating, organizing, and adaptive functions, the conditions under which emotions become maladaptive warrant further research. Extreme anger and fear can bring about large changes in the activities of internal organs innervated by the autonomic nervous system. When such arousal repeatedly involves the sympathetic nervous system and the hormones of the medulla of the adrenal gland, the individual may develop resistance to mental and physical disorders. When there is repeated arousal involving the sympathetic nervous system and the hormones of the cortex of the adrenal gland, the individual may experience adverse effects.

Problems of adaptation and mental health can also be conceived as attributable not to the emotions but to the

way a person thinks and acts. For example, if a person decides to break a moral code and consequently feels guilty, the guilt may be adaptive in that it can provide motivation for making amends. In this framework psychological problems or disorders arise because the individual fails to respond appropriately to the emotion's motivational cues while the emotion is still at low or moderate intensity.

THE REGULATION OF EMOTIONS

Several beliefs and attitudes have contributed to the idea that emotions should be brought under rather tight control. Historically, some religious and philosophical literature has treated human passion, a concept which included emotions, as an evil force that could contaminate or even destroy the mind or soul. In this tradition passions became associated with sin and wrongdoing, and their rigorous control was thus a sign of goodness. Even in this tradition, however, some negative emotions were exempt from tight control—guilt as a result of wrongdoing and righteous indignation toward moral transgressions.

Changing views of emotion regulation. Traditionally, scientists have given far more attention to negative emotions and their control than to positive ones. The focus on negative emotions has continued among clinical psychologists and psychiatrists, who are concerned with relieving depression and anxiety. However, as parents have long recognized, there is also a need to regulate positive emotions when, for example, children are having fun at someone else's expense or while neglecting chores and homework.

Developmental processes in emotion regulation. Of central importance in emotion regulation are developmental processes that enable children, as they mature, to exercise an increasingly greater control over affective responses. For example, before an infant can regulate the innate affective behaviour patterns elicited by acute pain, maturation of neural inhibitory mechanisms is required. Further control is realized through techniques that result from cognitive development and socialization, processes involving both maturation and learning.

Regulation in infants

In a study of responses of two- to 19-month-old infants to the pain of diphtheria-tetanus-pertussis (DTP) inoculation, it was found that the physical distress expression occurred as the initial response in all infants at the ages of two, four, and seven months (the ages at which the first three DTPs were administered). The physical distress expression is an all-out emergency response, a cry for help that dominates the physical and mental capacities of the infant. Beginning at the age of four months and accelerating rapidly between seven and 19 months, the infants became capable of greatly reducing the duration of the physical distress expression. As the duration of the physical distress expression decreased, that for anger expression increased. By 19 months of age, 25 percent of the infants were able to inhibit the distress expression completely. It was inferred that these developmental changes are adaptive for the relatively more capable toddler: whereas the physical distress expression in the younger subjects is all-consuming, anger mobilizes energy for defense or escape.

Other factors in emotion regulation. Several other factors are observable in emotion or mood regulation. First, there is neurochemical regulation by means of naturally occurring hormones and neurotransmitters. Regulation is also attained through psychoactive drugs, many of which were developed to control the prevalent psychological disorders of anxiety and depression. A substantial body of research has shown that anxiety and depression are associated with chemical imbalances in the brain and nervous system. Psychoactive drugs help to correct these imbalances.

Neurochemical regulation

Socialization processes, especially child-rearing practices, influence emotion regulation. Attempts by parents, teachers, and other adults to control emotions may be aimed either at the level of expression or experience or both. Parents may try to control their child's anger expressions before they culminate in "temper tantrums." A father may try to control his son's expressions of fear of bodily injury because he anticipates the shame of his son being seen as a coward. In considering the net effect of socialization on emotion regulation, it is necessary to weigh the effects that the child's unique genetic makeup may contribute to the process.

Cognitive-social theories point to cognitive processes as means of controlling emotion. According to this approach, if it is possible for people to change the way they make appraisals and attributions about the nature and cause of events, their emotion experiences can be changed. This could be manifested, for instance, in a reduction in self-blame and an alteration in negative concepts and outlooks. That cognitive therapy and cognitive techniques for controlling depressive and aggressive behaviour have achieved some success is testimony to the validity of the idea of cognitive control of emotion. That they sometimes fail indicates that it is no panacea and that other factors may be necessary for emotion regulation. As discussed above, theory and empirical data support the notion that expressive behaviour, which is under voluntary control, can be used to regulate emotions.

EMOTIONS, TEMPERAMENT, AND PERSONALITY

Most theorists agree that emotion thresholds and emotion responsiveness are part of the infrastructure of temperament and personality. There has, however, been little empirical research on the relations among measures of emotions, dimensions of temperament, and personality traits.

Emotions and temperament. Most theories of temperament define at least one dimension of temperament in terms of emotion. Two theories maintain that negative emotions form the core of one of the basic and stable dimensions of temperament. Another suggests that each of the dimensions of temperament is rooted in a particular discrete emotion and that these dimensions form the emotional substrate of personality characteristics. For example, proneness to anger would influence the development of aggressiveness, and the emotion of interest would account for the temperament trait of persistence.

Dimensions of temperament

Emotions and personality. A number of major personality theories, such as theories of temperament, identify dimensions or traits of personality in terms of emotions. For example, the German-born British psychologist Hans J. Eysenck has proposed three fundamental dimensions of personality: extroversion–introversion, neuroticism, and psychoticism. Extroversion–introversion includes the trait of sociability, which can also be related to emotion (*e.g.*, interest, as expressed toward people, versus shyness). Neuroticism includes emotionality defined, as in temperament theory, as nonspecific negative emotional responsiveness. Psychoticism may represent emotions gone awry or the absence of emotions appropriate to the circumstances.

Eysenck's fundamental dimensions

Several studies have shown that measures of positive emotionality and negative emotionality are independent, are not inversely related, and have stability over time. Further, it has been shown that positive and negative emotionality have different relations with symptoms of psychological disorders. For example, negative emotionality correlates positively with panic attack, panic-associated symptoms and obsessive-compulsive symptoms; that is, the higher the degree of negative emotion, the more likely that the attack or symptoms will occur. Conversely, positive emotionality correlates negatively with these phenomena. Although several of the same negative emotions characterize both the anxiety and depressive disorders, a lack of positive emotion experiences is more characteristic of depression than of anxiety.

Continuity of emotion expressiveness. Some studies have shown that specific emotions, identified in terms of expressive behaviour and physiological functions, have stability. One study showed that a child's expression of positive and negative emotion was consistent during the first two years of life. Other studies have shown stability of wariness or fear responses, indicating that a child who is fearful at one age is likely to be fearful in comparable situations at a later age. In a study of infants' responses to the pain of DTP inoculation, it was found that the child's anger expression indexes at ages two, four, and six months accurately predicted his or her anger expression in the inoculations at 19 months of age. Similar results were obtained for the sadness expression.

A study of mother–infant interaction and separation found that infants' expression at three to six months of age were accurate predictors of infant emotion expressive patterns at nine to 12 months of age. Emotion expression patterns have also shown continuity from 13 to 18 months of age during brief mother–infant separation.

Conclusion

The emotions are central to the issues of modern times, but perhaps they have been critical to the issues of every era. Poets, prophets, and philosophers of all ages have recognized the significance of emotions in individual life and human affairs, and the meaning of a specific emotion, at least in the context of verbal expression, seems to be timeless. Although art, literature, and philosophy have contributed to the understanding of emotion experiences throughout the ages, modern science has provided a substantial increase in the knowledge of the neurophysiological basis of emotions and their structure and functions.

Research in neuroscience and developmental psychology suggests that emotions can be activated automatically and unconsciously in subcortical pathways. This suggests that humans often experience emotions without reasoning why. Such precognitive information processing may be continuous, and the resulting emotion states may influence the many perceptual-cognitive and behavioral processes (such as perceiving, thinking, judging, remembering, imagining, and coping) that activate emotions through pathways involving the neocortex.

The two recognized types of emotion activation have important implications for the role of emotions in cognition and action. Subcortical, automatic information processing may provide the primitive data for immediate emotional response, whereas higher-order cognitive information processing involving the neocortex yields the evaluations and attributions necessary for the appropriate emotions and coping strategy in a complex situation.

Biosocial and constructivist theories agree that perception, thought, imagery, and memory are important causes of emotions. They also agree that once emotion is activated, emotion and cognition influence each other. How people feel affects what they perceive, think, and do, and vice versa.

Emotions have physiological, expressive, and experiential components, and each component can be studied in terms of its structure and functions. The physiological component influences the intensity and duration of felt emotion, expressions serve communicative and sociomotivational functions, and emotion experiences (feeling states) influence cognition and action.

Research has shown that certain emotion expressions are innate and universal and have significant functions in infant development and in infant–parent relations and that there are stable individual differences in emotion expressiveness. Emotion states influence what people perceive, learn, and remember, and they are involved in the development of empathic, altruistic, and moral behaviour and in basic personality traits.

BIBLIOGRAPHY. Studies of philosophical and cultural views on emotion include JAMES HILLMAN, *Emotion: A Comprehensive Phenomenology of Theories and Their Meanings for Therapy* (1960), a contemporary philosopher's explanation of emotions in terms of Aristotle's system of causes and a review of other approaches; AMÉLIE OKSENBERG RORTY (ed.), *Explaining Emotions* (1980), a collection of philosophical essays on the causes, meaning, and consequences of emotions; and ROM HARRÉ (ed.), *The Social Construction of Emotions* (1986), a collection of studies on the role of language and culture in the cognitive construction, *i.e.,* learning, of emotions.

The significance of emotions is the subject of many analyses, beginning with CHARLES DARWIN, *The Expression of the Emotions in Man and Animals* (1872, reprinted 1979), a classical work that placed human emotions in evolutionary perspective and presented the first evidence for their innateness and universality in human beings; CARROLL E. IZARD, *Human Emotions* (1977), a discussion of each of the fundamental emotions of human experience in terms of their unique organizing and motivational influence on cognition and action; SUSANNE K. LANGER, *Mind: An Essay on Human Feeling,* 3 vol. (1967–72), a philosopher's view of the significance of feelings in the evolution of human mentality; GEORGE MANDLER, *Mind and Body: Psychology of Emotion and Stress* (1984), a cognitive, or constructivist, view of the role of emotions in mental and bodily processes; ROBERT PLUTCHIK, *Emotion, a Psychoevolutionary Synthesis* (1980), a look at emotions in evolutionary perspective; and SILVAN S. TOMKINS, *Affect, Imagery, Consciousness,* vol. 1, *The Positive Affects* (1962), a brilliant essay on emotions as the primary motivational system of human beings.

The following works reflect some contemporary approaches to the study of emotions: MAGDA B. ARNOLD, *Emotion and Personality,* vol. 1, *Psychological Aspects* (1960), emphasizes the role of cognitive appraisal in emotion and sets the stage for later cognitive-social, or constructivist, theories of emotion; NICO H. FRIJDA, *The Emotions* (1986), is a comprehensive cognitive-social view of emotions; JOSEPH J. CAMPOS *et al.,* "Socioemotional Development," chapter 10 in MARSHALL M. HAITH and JOSEPH J. CAMPOS (eds.), *Infancy and Developmental Psychobiology,* 4th ed. (1983), pp. 783–915, provides a comprehensive review of theory and research on emotional development; ROBERT N. EMDE, THEODORE J. GAENSBAUER, and ROBERT J. HARMON, *Emotional Expression in Infancy: A Biobehavioral Study* (1976), is an influential contribution to the study of expressions; NATHAN A. FOX and RICHARD J. DAVIDSON (eds.), *The Psychobiology of Affective Development* (1984), presents a collection of reviews of theory and research papers on the biological aspects of emotional development; CARROLL E. IZARD, JEROME KAGAN, and ROBERT B. ZAJONC (eds.), *Emotions, Cognition, and Behavior* (1984), is a collection of research papers by leading psychologists on the relations between emotions, cognition, and actions; CARROLL E. IZARD and C.Z. MALATESTA, "Perspectives on Emotional Development I: Differential Emotions Theory of Early Emotional Development," chapter 9A in JOY DONIGER OSOFSKY (ed.), *Handbook of Infant Development,* 2nd ed. (1987), pp. 494–554, provides a detailed theory of emotional development and a review of related research; JOSEPH E. LEDOUX, "Emotion," chapter 10 in FRED PLUM (ed.), *Higher Functions of the Brain* (1987), pp. 419–59, in *Handbook of Physiology,* section 1, vol. 5, discusses brain mechanisms and neural pathways involved in the activation, expression, and experience of emotion; MICHAEL LEWIS and LINDA MICHALSON, *Children's Emotions and Moods: Developmental Theory and Measurement* (1983), explores a cognitive-social view of the development of emotions; PHOEBE C. ELLSWORTH and CRAIG A. SMITH, "From Appraisal to Emotion: Differences Among Unpleasant Feelings," *Motivation and Emotion,* 12(3):271–302 (September 1988), surveys research on the relations between appraisal processes and emotions and presents a new theory of cognition–emotion relations; H. HILL GOLDSMITH *et al.,* "What Is Temperament? Four Approaches," *Child Development,* 58(2):505–29 (April 1987), reviews theories of temperament with attention to temperament–emotion relations; ALICE M. ISEN, KIMBERLY A. DAUBMAN, and GARY P. NOWICKI, "Positive Affect Facilitates Creative Problem Solving," *Journal of Personality and Social Psychology,* 52(6):1122–31 (June 1987), exemplifies research showing how positive emotion facilitates creative thinking, empathy, and altruism; CARROLL E. IZARD, ELIZABETH A. HEMBREE, and ROBIN R. HUEBNER, "Infants' Emotion Expressions to Acute Pain: Developmental Change and Stability of Individual Differences," *Developmental Psychology,* 23(1):105–13 (January 1987), studies change and continuity in children's emotion expressions; WILLIAM JAMES, "What Is an Emotion?" *Mind,* 9:188–205 (1884), provides a classic definition of emotion that remains influential today; JEROME KAGAN, J. STEVEN REZNICK, and NANCY SNIDMAN, "Biological Bases of Childhood Shyness," *Science,* 240:167–71 (April 1988), summarizes a series of studies on biological bases and the continuity of shyness; and ROGER SPERRY, "Some Effects of Disconnecting the Cerebral Hemispheres," *Science,* 217:1223–26 (September 1982), discusses the effects of disconnecting cerebral hemispheres on mental and emotional experience.

(C.E.I.)

Encyclopaedias and Dictionaries

For more than 2,000 years encyclopaedias have existed as summaries of extant scholarship in forms comprehensible to their readers. The word encyclopaedia, of Greek origin (*enkyklopaideia*), at first meant a circle or a complete system of learning—that is, an all-around education. When Rabelais used the term in French for the first time in *Pantagruel* (chapter 20), he was still talking of education. It was Paul Scalich, a German writer and compiler, who was the first to use the word to describe a book in the title of his *Encyclopaedia; seu, orbis disciplinarum, tam sacrarum quam prophanum epistemon* . . . ("Encyclopaedia; or Knowledge of the World of Disciplines, Not Only Sacred but Profane . . . "), issued at Basel in 1559. The many encyclopaedias that had been published prior to this time either had been given fanciful titles (*Hortus deliciarum,* "Garden of Delights") or had been simply called "dictionary." The word dictionary has been widely used as a name for encyclopaedias, and Scalich's pioneer use of encyclopaedia did not find general acceptance until Denis Diderot made it fashionable with his historic French encyclopaedia, although cyclopaedia was then becoming fairly popular as an alternative term.

An outline of the scope and history of encyclopaedias is essentially a guide to the story of the development of scholarship, for encyclopaedias stand out as landmarks throughout the centuries, recording much of what was known at the time of publication. Many homes have no encyclopaedia at all, very few have more than one, yet in the past 2,000 years at least 2,000 encyclopaedias have been issued in various parts of the world, and some of these have had many editions. No library has copies of them all; if it were possible to collect them they would occupy some two miles of shelf space. But they are worth preserving—even those that appear to be hopelessly out-of-date—for they contain many contributions by a large number of the world's leaders and scholars.

"Dictionary" is used to describe a wide variety of reference works. Basically, a dictionary lists a set of words with information about them. The list may attempt to be a complete inventory of a language or may be only a small segment of it. A short list, sometimes at the back of a book, is often called a glossary. When a word list is an index to a limited body of writing, with references to each passage, it is called a concordance. Theoretically, a good dictionary could be compiled by organizing into one list a large number of concordances. A word list that consists of geographic names only is called a gazetteer.

The word lexicon designates a wordbook, but it also has a special abstract meaning among linguists, referring to the body of separable structural units of which the language is made up. In this sense, a preliterate culture has a lexicon long before its units are written in a dictionary. Scholars in England sometimes use "lexis" to designate this lexical element of language.

The compilation of a dictionary is lexicography; lexicology is a branch of linguistics in which, with the utmost scientific rigour, the theories that lexicographers use in the solution of their problems are developed.

The common phrase "dictionary order" takes for granted that the alphabetical order will be followed, and yet the alphabetical order has been called a tyranny that makes dictionaries less useful than they might be if compiled in some other order. The assembling of words into groups related by some principle, as by their meanings, can be done, and such a work is often called a thesaurus or synonymy. Such works, however, need an index for ease of reference, and it is unlikely that alphabetical order will be superseded except in specialized works. A monolingual dictionary has both the word list and the explanations in the same language, whereas bilingual or multilingual (polyglot) dictionaries have the explanations in another language or different languages. The word dictionary is also extended, in a loose sense, to reference books with entries in alphabetical order, such as a dictionary of biography, a dictionary of heraldry, or a dictionary of plastics.

For coverage of related topics in the *Macropædia* and the *Micropædia,* see the *Propædia,* section 735, and the *Index.*

The article is divided into the following sections:

ENCYCLOPAEDIAS

The meaning of the word encyclopaedia has changed considerably during its long history. Today most people think of an encyclopaedia as a multivolume compendium of all available knowledge, complete with maps and a detailed index, as well as numerous adjuncts such as bibliographies, illustrations, lists of abbreviations and foreign expressions, gazetteers, and so on. They expect it to include biographies of the great men and women of the present as well as those of the past, and they take it for granted that the alphabetically arranged contents will have been written in their own language by many people and will have been edited by a highly skilled and scholarly staff; nevertheless, not one of these ingredients has remained the same throughout the ages. Encyclopaedias have come in all sizes from a single 200-page volume written by one man to giant sets of 100 volumes or more. The degree of coverage of knowledge has varied according to the time and country of publication. Illustrations, atlases, and bibliographies have been omitted from many encyclopaedias, and for a long time it was not thought fitting to include biographies of living persons. Indexes are a late addition, and most of the early ones were useless. Alphabetical arrangement was as strongly opposed as the use of any language but Latin, at least in the first 1,000 years of publication in the West, and skilled group editorship has a history of scarcely 200 years.

In this article the word encyclopaedia has been taken to include not only the great general encyclopaedias of the past and the present but all types of works that claim to provide in an orderly arrangement the essence of "all that is known" on a subject or a group of subjects. This includes dictionaries of philosophy and of American history as well as volumes such as *The World Almanac and Book of Facts,* which is really a kind of encyclopaedia of current information.

The nature of encyclopaedias

In the *Speculum majus* ("The Greater Mirror"; completed 1244), one of the most important of all encyclopaedias, the French medieval scholar Vincent of Beauvais maintained not only that his work should be perused but that the ideas it recorded should be taken to heart and imitated. Alluding to a secondary sense of the word *speculum* ("mirror"), he implied that his book showed the world what it is and what it should become. This theme, that encyclopaedias can contribute significantly to the improvement of mankind, recurs constantly throughout their long history. A Catalan ecclesiastic and scholastic philosopher, Ramon Llull, regarded the 13th-century encyclopaedias, together with language and grammar, as instruments for the pursuit of truth. Domenico Bandini, an Italian humanist, planned his *Fons memorabilium universi* ("The Source of Noteworthy Facts of the Universe") at the beginning of the 15th century to provide accurate information on any subject to educated men who lacked books and to give edifying lessons to guide them in their lives. Francis Bacon believed that the intellect of the 17th-century individual could be refined by contact with the intellect of the ideal man. Another Englishman, the poet and critic Samuel Taylor Coleridge, was well aware of this point of view and said in his "Preliminary Treatise on Method" (1817) that in the *Encyclopædia Metropolitana,* which he was proposing to create, "our great objects are to exhibit the Arts and Sciences in their Philosophical harmony; to teach Philosophy in union with Morals; and to sustain Morality by Revealed Religion." He added that he intended to convey methodically "the pure and unsophisticated knowledge of the past . . . to aid the progress of the future." The Society for the Diffusion of Useful Knowledge declared in *The Penny Cyclopædia* (1833–43) that, although most encyclopaedias attempted to form systems of knowledge, their own would in addition endeavour to "give such general views of all great branches of knowledge, as may help to the formation of just ideas on their extent and relative importance, and to point out the best sources of complete information."

In *De disciplinis* ("On the Disciplines"; 1531) the Spanish humanist Juan Luis Vives emphasized the encyclopaedia's role in the pursuit of truth. In Germany of the early 19th century the encyclopaedia was expected to provide the right or necessary knowledge for good society. Probably the boldest claim was that of Alexander Aitchison, who said that his new *Encyclopædia Perthensis* (1796–1806) was intended to supersede the use of all other English books of reference.

Greek and Roman concepts

All these ideas were a far cry from the Greek concept, deriving from Plato, that in order to think better it is necessary to know all, and from the Roman attitude of the advisability of acquiring all useful knowledge in order to carry out one's tasks in life competently. The present concept of the encyclopaedia as an essential starting point from which one can embark on a voyage of discovery, or as a point of basic reference on which one can always rely, is little more than two centuries old.

The prose form has usually been accepted as the only suitable vehicle for the presentation of the text of an encyclopaedia, though *L'Image du monde* ("The Image of the World"; 1245?)—attributed by some to Gautier de Metz, a French poet and priest, and by others to a Flemish theologian, Gossuin—was written in French octosyllabic verse. It has also been generally accepted that an encyclopaedia should adopt a straightforward, factual approach. Even so, the Spanish writer Alfonso de la Torre, in his *Visio delectable* ("Delightful Vision"; *c.* 1484), adopted the allegorical approach of a child receiving instruction from a series of maidens named Grammar, Logic, Rhetoric, and so on.

The alphabetically arranged encyclopaedia has a history of less than 1,000 years, most of the encyclopaedias issued before the introduction of printing into Europe having been arranged in a methodical or classified form. The early compilers of encyclopaedias held, as Coleridge was to hold, that "to call a huge unconnected miscellany of the *omne scibile,* in an arrangement determined by the accident of initial letters, an encyclopaedia, is the impudent ignorance of your Presbyterian bookmakers!" Today several encyclopaedias still retain the classified form of arrangement.

There has never been any general agreement on the way in which the contents of an encyclopaedia should be arranged. In Roman times the approach was usually practical, with everyday topics such as astronomy and geography coming first, while the fine arts were relegated to the end of the work. The Roman statesman and writer Cassiodorus, however, in his 6th-century *Institutiones,* began with the Scriptures and the church and gave only brief attention to such subjects as arithmetic and geometry. St. Isidore of Seville, educated in the classical tradition, redressed the balance in the next century in his *Etymologiarum sive originum libri XX* ("Twenty Books on Origins, or Etymologies"), commonly called *Etymologiae,* giving pride of place to the liberal arts and medicine, the Bible and the church coming later, but still preceding such subjects as agriculture and warfare, shipping and furniture. The earliest recorded Arabic encyclopaedia, compiled by the Arab philologist and historian Ibn Qutayba, had a completely different approach, beginning with power, war, and nobility, and ending with food and women. A later Persian encyclopaedia, compiled in 975–997 by the Persian scholar and statesman al-Khwārizmī, started with jurisprudence and scholastic philosophy, the more practical matters of medicine, geometry, and mechanics being relegated to a second group labelled "foreign knowledge." The general trend in classification in the Middle Ages is exemplified by Vincent of Beauvais's *Speculum majus,* which was arranged in three sections: "Naturale"—God, the creation, man; "Doctrinale"—language, ethics, crafts, medicine; "Historiale"—world history. The encyclopaedists were, however, still uncertain

of the logical sequence of subjects, and although there were many who started with theological matters, there were just as many who preferred to put practical topics first.

A turning point came with Francis Bacon's plan for his uncompleted *Instauratio magna* ("Great Instauration"; 1620) in which he eschewed the endless controversies in favour of a three-section structure, including "External Nature" (covering such topics as astronomy, meteorology, geography, and species of minerals, vegetables, and animals), "Man" (covering anatomy, physiology, structure and powers, and actions), and "Man's Action on Nature" (including medicine, chemistry, the visual arts, the senses, the emotions, the intellectual faculties, architecture, transport, printing, agriculture, navigation, arithmetic, and numerous other subjects).

Various arrangements of the contents of encyclopaedias

In his plan Bacon had achieved more than a thoroughly scientific and acceptable arrangement of the contents of an encyclopaedia; he had ensured that the encyclopaedists would have a comprehensive outline of the scope of human knowledge that would operate as a checklist to prevent the omission of whole fields of human thought and endeavour. Bacon so profoundly altered the editorial policy of encyclopaedists that even 130 years later Diderot gratefully acknowledged his debt in the prospectus (1750) of the *Encyclopédie.* Because every later encyclopaedia was influenced by Diderot's work, the guidance of Bacon still plays its part today.

By courtesy of the trustees of the British Museum; photograph, R.B. Fleming & Co.

Engraved title page from the first edition of Francis Bacon's *Instauratio magna,* published in London, 1620.

Coleridge, who was very much impressed by Bacon's scheme, in 1817 drew up a rather different table of arrangement for the *Encyclopaedia Metropolitana.* It comprised five main classes: Pure Sciences—Formal (philology, logic, mathematics) and Real (metaphysics, morals, theology); Mixed and Applied Sciences—Mixed (mechanics, hydrostatics, pneumatics, optics, astronomy) and Applied (experimental philosophy, the fine arts, the useful arts, natural history, application of natural history); Biographical and Historical, chronologically arranged; Miscellaneous and Lexicographical, a gazetteer, and a philosophical and etymological lexicon. The fifth class was to be an analytical index.

Influence of Coleridge's classification

Although Coleridge's classification was altered by the publisher, and although the *Metropolitana* was an impressive failure, the ideas for it had a lasting influence. Even though nearly all encyclopaedias today are arranged alphabetically, the classifications of Bacon and Coleridge still enable editors to plan their work with regard to an assumed hierarchy of the various branches of human knowledge.

The concept of alphabetical order was well known to both the Greeks and Romans, but the latter made little use of it. Neither the Greeks nor the Romans employed it for encyclopaedia arrangement, with the exception of Sextus Pompeius Festus in his 2nd-century *De verborum significatu* ("On the Meaning of Words"). St. Isidore's encyclopaedia was classified, but it included an alphabetically arranged etymological dictionary. The 10th- or 11th-century encyclopaedic dictionary known as *Suidas* was the first such work to be completely arranged alphabetically, but it had no influence on succeeding encyclopaedias, although glossaries, when included, were so arranged. Bandini's *Fons memorabilium universi* ("The Source of Noteworthy Facts of the Universe"), though classified, used separate alphabetical orders for more than a quarter of its sections, and the Italian Domenico Nani Mirabelli's *Polyanthea nova* ("The New Polyanthea"; 1503) was arranged in one alphabetical sequence. These were rare exceptions, however; the real breakthrough came only with the considerable number of encyclopaedic Latin-language dictionaries that appeared in the early 16th century, the best known of which is a series of publications by the French printer Charles Estienne. The last of the great Latin-language encyclopaedias arranged in alphabetical order was *Encyclopaedia* (1630) by the German Protestant theologian and philosopher Johann Heinrich Alsted. The publication of *Le Grand Dictionnaire historique* ("The Great Historical Dictionary"; 1674) of Louis Moréri, a French Roman Catholic priest and scholar, confirmed public preference both for the vernacular and the alphabetically arranged encyclopaedia; this choice was emphasized by the success of the posthumous *Dictionnaire universel* (1690) by the French lexicographer Antoine Furetière.

From time to time important attempts have been made to reestablish the idea of the superiority of the classified encyclopaedia. Coleridge saw the encyclopaedia as a vehicle for enabling man to think methodically. He felt that his philosophical arrangement would "present the circle of knowledge in its harmony" and give a "unity of design and of elucidation." He did agree that his appended gazetteer and English dictionary would best be arranged alphabetically for ease of reference. By then, however, alphabetical arrangement had too strong a hold, and it was not until 1935 that a new major classified encyclopaedia began to appear—the *Encyclopédie française* ("French Encyclopaedia"), founded by Anatole de Monzie. The Dutch *Eerste nederlandse systematisch ingerichte encyclopaedie* ("First Dutch Systematic and Comprehensive Encyclopaedia"; 1946–52) has a classification that is in almost reverse order of that of the *Encyclopédie française,* but it is clear that behind both works lies a philosophical concept of the order and main divisions of knowledge that is influenced by both Bacon and Coleridge. The Spanish *Enciclopedia labor* (1955–60) and the *Oxford Junior Encyclopaedia* (1948–56) follow systems of arrangement that are closer to the French than to the Dutch example.

Role of the trivium and quadrivium

From earliest times it had been held that the trivium (grammar, logic, rhetoric) and the quadrivium (geometry, arithmetic, astronomy, music) were essential ingredients in any encyclopaedia. Even as late as 1435 Alfonso de la Torre began his *Visiõ delectable* in almost that exact order, and only when he had laid these foundations did he proceed to the problems of science, philosophy, theology, law, and politics. Thus the seven liberal arts were regarded by the early encyclopaedists as the very mathematics of human knowledge, without a knowledge of which it would be foolish to proceed. This idea survived to a certain extent in Coleridge's classification; he stated that grammar and logic provide the rules of speech and reasoning, while

mathematics opens mankind to truths that are applicable to external existence.

When Louis Shores became editor in chief of *Collier's Encyclopedia* in 1962, he said that he considered the encyclopaedia to be "one of the few generalizing influences in a world of overspecialization. It serves to recall that knowledge has unity." This echoes the view of the English novelist H.G. Wells, that the encyclopaedia should not be "a miscellany, but a concentration, a clarification and a synthesis." The Austrian sociologist Otto Neurath in the same year suggested that a proposed new international encyclopaedia of unified science should be constructed like an onion, the different layers enclosing the "heart"—comprising in this case the foundations of the unity of science.

Even a brief survey of contemporary encyclopaedia publishing is enough to make clear that, as the trivium and quadrivium and the topically classified encyclopaedias that they influenced receded further and further into history, there arose a number of modern encyclopaedists concerned with the importance of making a restatement of the unity of knowledge and of the consequent interdependence of its parts. Though most encyclopaedists were willing to accept the essential reference-book function of encyclopaedias and the role of an alphabetical organization in carrying out that function, they became increasingly disturbed about the emphasis on the fragmentation of knowledge that such a function and such an organization encouraged. A number looked for ways of enhancing the educational function of encyclopaedias by reclaiming for them some of the values of the classified or topical organizations of earlier history.

Notable among the results of such activities was the 15th edition of *Encyclopædia Britannica* (1974), which was designed in large part to enhance the role of an encyclopaedia in education and understanding without detracting from its role as a reference book. Its three parts (*Propædia,* or *Outline of Knowledge; Micropædia,* or *Ready Reference and Index;* and *Macropædia,* or *Knowledge in Depth*) represented an effort to design an entire set on the understanding that there is a circle of learning and that an encyclopaedia's short informational articles on the details of matter within that circle as well as its long articles on general topics must all be planned and prepared in such a way as to reflect their relation to one another and to the whole of knowledge. The *Propædia* specifically was a reader's version of the circle of learning on which the set had been based and was organized in such a way that a reader might reassemble in meaningful ways material that the accident of alphabetization had dispersed.

Encyclopaedias in general

THE ROLE OF ENCYCLOPAEDIAS

Of the various types of reference works—who's whos, dictionaries, atlases, gazetteers, directories, and so forth—the encyclopaedia is the only one that can be termed self-contained. Each of the others conveys some information concerning every item it deals with; only the encyclopaedia attempts to provide coverage over the whole range of knowledge, and only the encyclopaedia attempts to offer a comprehensive summary of what is known of each topic considered. To this end it employs many features that can help in its task, including pictures, maps, diagrams, charts, and statistical tables. It also frequently incorporates other types of reference works. Several modern encyclopaedias, from the time of Abraham Rees's *New Cyclopaedia* (1802–20) and the *Encyclopédie méthodique* ("Systematic Encyclopaedia"; 1782–1832) onward, have included a world atlas and a gazetteer, and language dictionaries have been an intermittent feature of encyclopaedias for most of their history.

Treatment of biographical material

Most modern encyclopaedias since the *Universal-Lexicon* (1732–50) of the Leipzig bookseller Johann Heinrich Zedler have included biographical material concerning living persons, though the first edition of *Encyclopædia Britannica* (1768–71) had no biographical material at all. In their treatment of this kind of information they differ, however, from the form of reference work that limits itself to the provision of salient facts without comment. Similarly, with dictionary material, some encyclopaedias—such as the great Spanish "Espasa" (1905 to date)—provide foreign-language equivalents as well.

An English lexicographer, Henry Watson Fowler, wrote in the preface to the first edition (1911) of *The Concise Oxford Dictionary of Current English* that a dictionary is concerned with the uses of words and phrases and with giving information about the things for which they stand only so far as current use of the words depends upon knowledge of those things. The emphasis in an encyclopaedia is much more on the nature of the things for which the words and phrases stand. Thus the encyclopaedic dictionary, whose history extends as far back as the 10th- or 11th-century *Suidas,* forms a convenient bridge between the dictionary and the encyclopaedia, in that it combines the essential features of both, embellishing them where necessary with pictures or diagrams, at the same time that it reduces most entries to a few lines that can provide a rapid but accurate introduction to the subject.

Interrelations. An encyclopaedia does not come into being by itself. Each new work builds on the experience and contents of its predecessors. In many cases the debt is acknowledged: the German publisher Friedrich Arnold Brockhaus bought up the bankrupt encyclopaedia of Gotthelf Renatus Löbel in 1808 and converted it into his famous *Conversations-Lexikon;* but Jesuits adapted Antoine Furetière's *Dictionnaire universel* without acknowledgment in their *Dictionnaire de Trévoux* (1704). Classical writers made many references to their predecessors' efforts and often incorporated whole passages from other encyclopaedias. Of all the many examples, the *Cyclopaedia* (1728) of the English encyclopaedist Ephraim Chambers has been outstanding in its influence, for Diderot's and Rees's encyclopaedias would have been very different if Chambers had not demonstrated what a modern encyclopaedia could be. In turn, the publication of *Encyclopædia Britannica* was stimulated by the issue of the *Encyclopédie.* Almost every subsequent move in encyclopaedia making is thus directly traceable to Chambers' pioneer work.

The intended audience

Readership. Encyclopaedia makers have usually envisaged the particular public they addressed. Cassiodorus wrote for the "instruction of simple and unpolished brothers"; the Roman statesman Cato wrote for the guidance of his son; Gregor Reisch, prior of the Carthusian monastery of Freiburg, addressed himself to "Ingenuous Youth"; the Franciscan encyclopaedist Bartholomaeus Anglicus wrote for *ordinary* people; the German professor Johann Christoph Wagenseil wrote for children; and Herrad of Landsberg, abbess of Hohenburg, wrote for her nuns. *Encyclopædia Britannica* is designed for the use of the curious and intelligent layman. The editor of *The Columbia Encyclopedia* in 1935 tried to provide a work compact enough and simply enough written to serve as a guide to the "young Abraham Lincoln." The Jesuit Michael Pexenfelder (1670) made his intended audience clear enough by writing his *Apparatus Eruditionis* ("Apparatus of Learning") in the form of a series of conversations between teacher and pupil. St. Isidore addressed himself to the needs not only of his former pupils in the episcopal school but also to all the priests and monks for whom he was responsible. At the same time, he tried to provide the newly converted population of Spain with a national culture that would enable it to hold its own in the Byzantine world.

Contributors. In sympathy with many of their various ends, many scholars have contributed to encyclopaedias. Not all their contributions are known because, until recently, it was not the custom to sign articles. Even today there is what amounts to partial concealment in that articles are often initialled only, and, although a key is provided, few readers look up the writer's identity. It is known, however, that the English encyclopaedist John Harris enlisted the help of such scientists as John Ray and Sir Isaac Newton for his *Lexicon Technicum* (1704) and that Rees's *New Cyclopaedia* (1802–20) included articles on music by the English organist and music historian Charles Burney and on botany by the English botanist Sir J.E. Smith. Illustrious Frenchmen such as Voltaire, Rousseau, Condorcet, Montesquieu, and Georges Boulanger contributed to the *Encyclopédie;* the writer and statesman Thomas

Macaulay, the Russian-born jurist and medieval historian Sir Paul Vinogradoff, and the Czech statesman Tomáš Masaryk to the *Britannica;* the Scottish physicist Sir David Brewster and the Danish physicist Hans Christian Ørsted to *The Edinburgh Encyclopaedia* (1808–30); the English astronomer Sir William Herschel and the English mathematician and mechanical genius Charles Babbage to the *Metropolitana;* the Russian Communist leader Lenin to the "Granat" encyclopaedia; and the dictator Benito Mussolini to the *Enciclopedia italiana.*

Language. The language of Western encyclopaedias was almost exclusively Latin up to the time of the first printed works. As with most scholarly writings, the use of Latin was advantageous because it made works available internationally on a wide scale and thus promoted unlimited sharing of information. On the other hand, it made the contents of encyclopaedias inaccessible to the great majority of people. Consequently, there was from the early days on a movement to translate the more important encyclopaedias into various vernaculars. Honorius Inclusus' *Imago mundi* ("Image of the World"; *c.* 1122) was rendered into French, Italian, and Spanish; Bartholomaeus Anglicus' *De proprietatibus rerum* ("On the Characteristics of Things"; 1220–40) into English; the Dominican friar Thomas de Cantimpré's *De natura rerum* ("On the Nature of Things"; *c.* 1228–44) into Flemish and German; and Vincent of Beauvais's *Speculum majus* ("The Greater Mirror") into French, Spanish, German, Dutch, and Catalan. In later years the more successful encyclopaedias were translated from one vernacular into another. Moréri's encyclopaedia, *Le Grand Dictionnaire historique,* was translated into both English and German. The German *Brockhaus* appeared in a Russian translation (1890–1907), and the French *Petit Larousse* had several foreign-language editions. Nevertheless, an encyclopaedia, however successful in its own country, may find acceptance in another country far from easy, because each nation appears to have its own very individual concept of what an encyclopaedia should comprise.

Translations of Latin encyclopaedias

The contemporary world. Encyclopaedias have often reflected fairly accurately the civilization in which they appeared; that this was deliberate is shown by the frequency with which the earlier compilers included such words as *speculum* ("mirror"), *imago* ("image"), and so forth in their titles. Thus as early as the 2nd century the Greek sophist Julius Pollux was already defining current technical terms in his *Onomastikon.* In the 13th century Vincent of Beauvais quoted the ideas of both pagan and Christian philosophers freely and without differentiation, for their statements often agreed on questions of morals. In doing so, he reflected the rapidly widening horizons of a period that saw the founding of so many universities. Bartholomaeus Anglicus devoted a considerable part of his work to psychology and medicine. "Theophilus" (thought to be Roger of Helmarshausen, a Benedictine monk) as early as the 12th century gave a clear and practical account in his *De diversis artibus* ("On Diverse Arts") of contemporary processes used in painting, glassmaking and decoration, metalworking, bone carving, and the working of precious stones, even listing the necessary tools and conditions for successful operations. Pierre Bayle, a French philosopher and critic, showed in his *Dictionnaire historique et critique* ("Historical and Critical Dictionary"; 1697) how the scientific renaissance of the previous 40 years had revolutionized contemporary thought. To every detail he applied a mercilessly scientific and inquiring mind that challenged the assumptions and blind reverence for authority that had characterized most of his predecessors.

At that point in history, much attention was being paid to practical matters: the statesman Jean-Baptiste Colbert himself directed the French Académie des Sciences (1675) to produce a work that eventually appeared as the *Description et perfection des arts et métiers* ("Description and Perfection of the Arts and Crafts"; 1761). The German Meyer's *Grosses Conversations-Lexicon* from the first edition (1840–55) onward paid particular attention to scientific and technical developments, and the *Encyclopedia Americana,* aided by the *Scientific American,* strengthened its coverage in this area from 1911 onward. In its very first edition the *Encyclopædia Britannica* included lengthy articles containing detailed instructions on such topics as surgery, bookkeeping, and many aspects of farming. Similarly, Abraham Rees had been including articles on subjects such as candle making and coachbuilding earlier in the century. The outstanding example of a completely contemporary encyclopaedia was, of course, the *Encyclopédie,* in which the philosopher Denis Diderot and the mathematician and philosopher Jean Le Rond d'Alembert and their friends set out to reject much of the heritage of the past in favour of the scientific discoveries and the more advanced thought of their own age. Their decision in this respect was both intellectually and commercially successful; since that time every edition of any good encyclopaedia has the additional merit of being a valuable source for the thought and attitudes of the world for which it was published.

Treatment of practical matters in encyclopaedias

Encyclopaedias and politics. All great encyclopaedia makers have tried to be truthful and to present a balanced picture of civilization as they knew it, although it is probable that no encyclopaedia is totally unbiassed. A great encyclopaedia is inevitably a sign of national maturity and, as such, will pay tribute to the ideals of its country and its times. The first Hungarian encyclopaedia, János Apáczai Csere's *Magyar encyclopaedia* (1653–55), was mostly a summary of what was available in foreign works, but the *Révai nagy lexikona* ("Révai's Great Lexicon"; 1911–35) was a handsome tribute to Hungary's emergence as a country in its own right, just as the *Enciklopedija Jugoslavije* (1955–) is a prestige work that does full justice to the advances made by Yugoslavia in the mid-20th century. The supreme example of an encyclopaedia that set out to present the best possible image of its people and the wealth and stature of their culture is undoubtedly the *Enciclopedia italiana* (1929–36). Mussolini's contribution of an article on Fascism indicates the extent to which the work might be regarded as an ideological tool, but, in fact, the bulk of its contents is admirably international and objective in approach. The various Soviet encyclopaedias already occupy many feet of shelf space, and the later editions each devote one complete volume to the Soviet Union in all its aspects. Though successive editions have been notable for the obvious political factors that have been responsible for the inclusion and exclusion of entries for famous nationals according to the current state of their acceptance or condemnation by the existing regime, many critics have felt that the newest edition, the first volume of which was issued in 1970, is maturer than any of the others in this regard.

Opposition to encyclopaedias

Diderot, the editor, and André-François Le Breton, the publisher, faced such opposition from both church and state in their publication of the *Encyclopédie* (1751–65) that many of the volumes were secretly printed, and the last 10 were issued with a false imprint. In the early part of the 19th century, *Brockhaus* was condemned by the Austrian censor, and in 1950 its 11th edition was branded as reactionary by the East German government. Nor was political censorship the only form of oppression in the world of encyclopaedias. Antoine Furetière, on issuing his prospectus (1675) for his *Dictionnaire universel,* found his privilege to publish cancelled by the French government at the request of the Académie Française, which accused him of plagiarizing its own dictionary. The Leipzig book trade, fearing that publication of Johann Heinrich Zedler's huge *Universal-Lexikon* (1732–50) might put them out of business, made such difficulties that Zedler thought it best to issue his work in Halle.

The reader's needs. People look to encyclopaedias to give them an adequate introduction to a topic that interests them. Many expect the encyclopaedia to omit nothing and to include consideration of all controversial aspects of a subject. Encyclopaedia makers of the past assumed that there was a large public willing to read through an entire encyclopaedia if it was not too large. In the 18th century, for example, there was a good market for pocket-size compendia for the traveller, or for the courtier to browse in as he waited for an audience. Thus, although most encyclopaedias are multivolume works, there are many small works ranging from the *Didascalion* of the scholastic

philosopher and mystic theologian Hugh of Saint-Victor (*c.* 1128), through Gregor Reisch's *Margarita philosophica* (1496) and the French writer Pons-Augustin Alletz' *Petite Encyclopédie* (1766), to C.T. Watkins' *Portable Cyclopaedia* (1817). The last was issued by a remarkable publisher, Sir Richard Phillips, who realized the great demand for pocket-size compendia and drove a thriving trade in issuing a number of these; he is thought to have written large sections of these himself.

Royalty and encyclopaedias. Most of the classic Chinese encyclopaedias owe their existence to the patronage of emperors. In the West, the Roman scholar Pliny dedicated his *Historia naturalis* ("Natural History") to the emperor Titus; Julius Pollux dedicated his *Onomastikon* to his former pupil, the Roman emperor Commodus, while the Byzantine philosopher and politician Michael Psellus dedicated his *De omnifaria doctrina* ("On All Sorts of Teaching") to his former pupil the emperor Michael VII Ducas, ruler of the Eastern Roman Empire. Gervase of Tilbury, an English ecclesiastic, compiled his *Otia imperialia* ("Imperial Pastimes") for the Holy Roman emperor Otto IV, and Alfonso de la Torre prepared his *Visiõ delectable* for Prince Carlos of Viana. St. Isidore dedicated his encyclopaedia to the Visigothic king Sisebut, and the French king Louis IX patronized Vincent of Beauvais's *Speculum majus.* Nor did kings eschew the work of compiling encyclopaedias. The emperor Constantine VII of the Eastern Roman Empire was responsible for a series of encyclopaedias, and Alfonso X of Spain organized the making of the *Grande e general estoria* ("Great and General History").

Editorial censorship

Contents and authority. The extent to which readers have been dependent on editorial decisions concerning not only what to include but also what to exclude has yet to be explored in detail. For example, Vincent of Beauvais rarely mentioned the pagan and Christian legends that were so popular in his day. The anonymous compiler of the scholarly *Compendium philosophiae* ("Compendium of Philosophy"; *c.* 1316) was careful to omit the credulous tales that appeared in contemporary bestiaries. For many centuries it was not considered right to include biographies of men and women who were still alive. And the early Romans, such as Cato the Censor, rejected much of Greek theoretical knowledge, regarding it as a dangerous foreign influence and believing with the Stoics that wisdom consisted in living according to nature's precepts.

Whatever the compiler did decide to include had a far-reaching influence. Pliny's vast *Historia naturalis* has survived intact because for so many centuries it symbolized human knowledge, and even the "old wives' tales" it injudiciously included were unquestioningly copied into many later encyclopaedias. The influence of St. Isidore's work can be traced in writings as late as Sir John Mandeville's travels (published in French between 1357 and 1371) and the English poet John Gower's 14th-century *Confessio amantis* ("A Lover's Confession"). Honorius' *Imago mundi* is known to have influenced some of the German medieval chronicles and the Norse saga of Olaf Tryggvason. The main source of classics such as the *Roman de la rose,* the Alexander romances, Archbishop Giovanni da Colonna's *Liber de viris illustribus* ("Book Concerning Illustrious Men"), and the recorded lives of the saints can be traced to the *Speculum majus.* The direct and indirect influence of the critical encyclopaedias of Pierre Bayle and Denis Diderot is, of course, incalculable.

EDITING AND PUBLISHING

The length of encyclopaedias and encyclopaedic articles. There always have been and there still are a number of successful one-volume encyclopaedias. Current outstanding examples include *The Columbia Encyclopedia,* the *Petit Larousse, Hutchinson's New Twentieth Century Encyclopedia,* and the *Random House Encyclopedia.* In the Random House set the contents were divided into two sections, a *Colorpedia,* composed of relatively lengthy articles dealing with broad topics, and an *Alphapedia,* composed of concise entries on very specific subjects. Some booksellers and publishers confirm that there is, however unreasonably, a certain amount of public prejudice against the single-volume form, and that most people prefer a multivolume work. Throughout the entire history of encyclopaedias there has been much variation in the number of volumes. Many of the Chinese encyclopaedias have been very much larger than any Western work. Pliny's *Historia naturalis* comprised about 2,500 chapters, Zedler's *Universal-Lexicon* was planned for 12 volumes and eventually filled 64; the publishers of the *Encyclopédie* were faced with a lawsuit (1768–78) for producing a 26-volume encyclopaedia instead of the 10 volumes they had promised; and Johann Samuel Ersch and Johann Gottfried Gruber's German *Allgemeine Encyclopädie* ("General Encyclopaedia") had already reached 167 volumes at the time of its discontinuance. Today, although the major Soviet encyclopaedia consists of more than 50 volumes, most encyclopaedias range between 20 and 30 volumes, occupying between three and four feet of shelf space. Thus the modern encyclopaedia appears smaller than its 19th-century counterpart, but, in fact, the content may be greater because the thick mat paper of Victorian times has been replaced by a thinner paper capable of reproducing coloured and black-and-white halftone illustrations with sharp definition.

Even more noticeable than variations in the number of volumes in encyclopaedias has been an even greater variation in the average lengths of articles within those volumes. The 11th edition of the *Encyclopædia Britannica* contained almost twice as many articles as the last significant edition before it, but it contained only 15 or 16 percent more words. The difference had to do with editorial considerations regarding the matter of fragmentation. Although most of the major encyclopaedias of the past had devoted considerable space to any topic of major importance, there was increasing recognition in the 19th century that an alternative method of treatment would be to break large subjects into their constituent subtopics for alphabetical distribution throughout the set. Those who favoured this more fragmented approach argued that by focussing on the smaller part of the whole, the editors could facilitate the user's search for specific information and that the liberal provision of cross-references would facilitate a recombination of the fragments by those interested in the bigger picture. Against this practice, it was argued that most cross-references are not followed up by most readers, that the shorter fragmented pieces work against a correct understanding of the larger subject, and that fragmentation inevitably involved a great amount of repetition of basic information throughout all of the related articles. Nevertheless, *Brockhaus, Meyer, Larousse,* and other encyclopaedias of the shorter entry type have had and continue to have a strong following.

Authorship by one individual

Authorship. The first encyclopaedia makers had no doubts concerning their ability to compile their works single-handedly. Cassiodorus, Honorius Inclusus (or Solitarius), and Vincent of Beauvais fully justified this attitude, though their task was largely that of the anthologist. Vincent and many other encyclopaedists employed both scribes and scholars to help them in their work, but once the encyclopaedia reached the stage of independent writing it was clear that the editorial task was going to become more complex. Even so, some of the later pocket encyclopaedias—such as the English bookseller John Dunton's mediocre *Ladies' Dictionary* (1694), *An Universal History of Arts and Sciences* (1745) by the French-born Englishman Chevalier Denis de Coëtlogon, and the popular *Allgemeines Lexicon* (1721) by the Prussian scholar Johann Theodor Jablonski—were substantially or almost wholly the work of a single author; such items are, however, negligible.

John Harris, an English theologian and scientist, may have been one of the first to enlist the aid of experts, such as the naturalist John Ray and Sir Isaac Newton, in compiling his *Lexicon Technicum* ("Technical Lexicon"; 1704). Johann Heinrich Zedler, in his *Universal-Lexicon* (1732–50), went further by enlisting the help of two general editors, supported by nine specialist editors, the result being a gigantic work of great accuracy. The French *Encyclopédie,* the largest encyclopaedia issued at that time, inevitably had many contributors, although the French writer Voltaire said that Diderot's collaborator, the

Chevalier Louis de Jaucourt (aided by secretaries), contributed about three-quarters of the articles in that work. The pattern for future encyclopaedias was established: for any substantial work it would be necessary not only to have contributions from the experts of the day, but it would also be essential to have subject editors who could supervise the coverage and content in each area of knowledge. With little alteration, this system prevails today.

Encyclopaedia adjuncts. The readers of modern encyclopaedias are rarely aware of the numerous aids that have been provided to make their search for information so easy and efficient. Only when recourse is had to one of the older encyclopaedias does the reader become conscious of the advances that have been made. In former days it was often difficult to distinguish between one article and the next, because distinctive headings or inset titles or the use of boldface was rare. Nor was the necessity for running titles or alphabetical notations at the head of the pages fully appreciated. Even more troublesome was the problem of the arrangement of entries for several persons of the same name; reference to the older encyclopaedias under such headings as "Henry," "Charles," "John," or "Louis"—names held by both princes and religious potentates—will show how little the art of acceptable arrangement was understood.

Arrangement of articles

Cross-references and bibliographies. Cross-references are an essential feature of the modern encyclopaedia; they date back at least as far as Bandini's *Fons memorabilium universi,* but it was Brockhaus who introduced an ingenious system of using arrows instead of the words "see also." *The Columbia Encyclopedia* achieves the same effect by printing in small capital letters the words under which additional information can be found. Other features of interest in the modern encyclopaedia include the devotion of each volume to a letter of the alphabet (*Compton's Encyclopedia* formerly followed this system), or the indication of the division between letters by thumb-indexing or by the insertion of a thicker sheet of distinctively coloured paper. In established encyclopaedias the bibliographies for individual articles are usually the result of careful editorial consultation with the writer and with librarians.

Indexes. Undoubtedly the major adjunct of the modern encyclopaedia is its index. As early as 1614 the bishop of Petina, Antonio Zara, included an index of a kind in his *Anatomia ingeniorum et scientiarum* ("Anatomy of Talents and Sciences"). A Greek professor at Basel, Johann Jacob Hoffman, added an index to his *Lexicon universale* of 1677; the *Encyclopédie* was completed by a two-volume "Table analytique et raisonnée" for the entire 33 volumes of text, supplements, and plates; and the *Britannica* included individual indexes to the lengthier articles in its 2nd edition (1778–84) and provided its first separate index volume for the 7th edition (1830–42). The nature of good indexing was still far from being fully understood, however, and it was only later in the 19th century that really good encyclopaedia indexes were prepared. In the 20th-century encyclopaedias that provide indexes—and there are many that do not—the reader is invariably advised to read the guides to their use because the index is a sophisticated tool that, by the aid of a few simple typographical devices and editorial conventions, is able to offer a wealth of information in one alphabetical sequence.

Illustrative material. The use of illustrations in encyclopaedias goes back almost certainly to St. Isidore's time. One of the most beautiful examples of an illustrated encyclopaedia was the abbess Herrad's 12th-century *Hortus deliciarum.* In many earlier encyclopaedias the illustrations were often more decorative than useful, but from the end of the 17th century the better encyclopaedias began to include engraved plates of great accuracy and some of great beauty. The *Encyclopédie* is particularly distinguished for its superb volume of plates—they have even been reprinted in the 20th century. In modern times the trend has been toward more lavish illustration of encyclopaedias, including elaborate coloured anatomical plates with superimposed layers, and specially inset small coloured halftones, as well as marginal line drawings. Since 1950 a form of encyclopaedia has begun to appear that comprises large numbers of coloured illustrations with a somewhat subordinated text.

Early illustrated encyclopaedias

The level of writing. The American editor Franklin H. Hooper, undaunted by his own lack of scholarship, took a notable part in ensuring that the articles of the 11th edition of *Encyclopædia Britannica* were kept within the mental range of the average man. The problem of the encyclopaedist has always been to strike the right mean between too learned and too simplified an approach. The Roman Cassiodorus wrote his encyclopaedia to provide a bridge between his unlettered monks and the scholarly books he had preserved for their use. Hugh of Saint-Victor, the theologian and philosopher, achieved one of the best approaches in his charming *Didascalion* (*c.* 1128), in which he used an elegant and simple style that everyone could appreciate. The abbess Herrad, knowing her audience, described in didactic fashion the history of the world (with emphasis on biblical stories) and its content, with commentaries and beautifully coloured miniatures designed to help and edify the simple nuns in her charge. The master of Dante, Brunetto Latini, wanted to reach the Italian cultured and mercantile classes with his *Li livres dou trêsor* ("Treasure Books"; *c.* 1264) and therefore used a concise and accurate style that evoked an immediate and general welcome. Gregor Reisch managed to cover the whole university course of the day in his very pleasing and brief *Margarita philosophica,* which correctly interpreted the taste of the younger generation at the end of the 15th century.

Until the 17th century there had been a great many encyclopaedias written by clerics for clerics, and further examples continued to be published. After that time, more popular works began to be published as well, particularly in France, where such palatable compilations as the Sieur Saunier's *Encyclopédie des beaux esprits* ("Encyclopaedia of Great Minds"; 1657) had an immediate success. The philosopher Pierre Bayle in his *Dictionnaire historique et critique* (1697) first introduced the lay reader to the necessity of reading more critically; in this his work constituted a forerunner of the *Encyclopédie,* with its challenges to so many undiscriminating assumptions about religion and politics, history and government. On the other hand, the contemporary *Dictionnaire universel* of the Jesuit fathers of Trévoux had a popularity among the orthodox that caused it to run through six editions and gradually to increase its size from three to eight volumes between 1704 and 1771.

Encyclopaedias for clerics and lay readers

Supplementary material. The idea of keeping encyclopaedias up-to-date by means of supplements, yearbooks, and so on, is well over 200 years old. In 1753 a two-volume supplement to the 7th edition of Ephraim Chambers' *Cyclopaedia* was compiled by George Lewis Scott, a tutor to the English royal family. Charles-Joseph Panckoucke, a publisher, issued a four-volume supplement to the *Encyclopédie* (1776–77), in spite of Diderot's refusal to edit it. The *Britannica* included a 200-page appendix in the last volume of the 2nd edition (1784) and issued a two-volume supplement to the 3rd edition (1801; reprinted 1803). *Brockhaus* broke new ground by issuing in monthly parts (1857–64) a yearbook to the 10th edition (1851–55), which, on the commencement of the issue of the 11th edition, changed its name to *Unsere Zeit* ("Our Times") and doubled its frequency (1865–74). In 1907 Larousse began publication of the *Larousse mensuel illustré* ("Monthly Illustrated Larousse"). *The New International Encyclopaedia* issued a yearbook from 1908 (retrospective to 1903), and the *Britannica* issued one yearbook in 1913 and recommenced with the *Britannica Book of the Year* in 1938. The publication of supplements has a much longer history in China, but the system on which the Chinese operated was very different from that of the West (see below). Nowadays yearbooks are a common feature of encyclopaedias of standing in the United States and other countries. In the main, they are more effective in recording the events and discoveries of each year than they are in keeping the main articles up-to-date, but they perform an essential duty in informing their readers of much that is not reported or that is only inadequately reported in the press; at the same time, they provide a more

reasoned assessment and perspective than the daily newspapers and the weekly commentaries can usually achieve.

Some of the leading encyclopaedias offer additional services that are not too widely known. The modern encyclopaedia is a complex work of reference, and the reader needs expert guidance if he is to get the best from its contents. To this end, small subject guides are sometimes issued, which, in narrative form, outline the whole field and bring each topic into perspective, drawing attention to the appropriate articles that will throw further light on the matter. As a means of self-education, some of these outlines maintain a high standard and are an invaluable adjunct to the encyclopaedia. Another supplementary feature offered by some established encyclopaedias is a research service through which purchasers are permitted to submit a limited number of questions about topics either not dealt with in the set or dealt with inadequately. Such services have been provided in a variety of ways. In some cases, frequently asked questions may be answered with previously prepared reports listed in the publisher's catalog; in others, questions are referred to a special office staff for answers culled from the publisher's own data bases; and in still others they may be referred to researchers stationed at selected specialized libraries. Though increasingly expensive to maintain, such services have been thought to be advantageous to the reader, who secures from them a research support not otherwise available, and advantageous to the editorial staff, who may derive from them useful information about the kinds of problems being experienced by those who use their encyclopaedia.

Other supplementary material sometimes issued by encyclopaedias ranges from 10-year illustrated surveys of events to sets of books that have had a major impact on mankind. Although few publishers include dictionaries as an integral part of their encyclopaedia, they frequently supply a well-known, independently compiled work as part of their service. It is an increasingly common custom, however, for a modern encyclopaedia to incorporate an atlas and a gazetteer, often in the last volume.

Problems of encyclopaedias. *Authorship.* In using a reputable encyclopaedia the reader is inclined to accept the authenticity of any article he happens to read. Subconsciously he is aware that the highly organized body of scholars listed at the beginning of the work must inevitably have ensured the severe scrutiny of all material. Nevertheless, in recent years editors of encyclopaedias have tended more and more to commission signed articles by well-known experts. One of the most famous examples of this was the *Britannica's* commissioning of articles for its 1922 supplement from some of the most famous men and women of the day: "Belgium" by the Belgian historian Henri Pirenne; "Anton Ivanovich Denikin" by the Russian-born jurist and historian Sir Paul Vinogradoff; "Drama" by St. John Ervine, the Irish playwright and novelist; "Czechoslovakia" by the Czech statesman Tomáš Masaryk; and "Russian Army" by Gen. Yuri Danilov. This created a new dimension in encyclopaedias, for it introduced a personal element on a scale previously seen only in the columns of the *Encyclopédie.*

Famous contributors

There is, in fact, a difference in the treatment of a subject written by a politician such as Masaryk and by an academic historian of distinction. Each writer has something important to offer, and the results will be very different.

Length restrictions. The restrictions imposed by the space available for any particular article are of great consequence. If an expert is asked to contribute an article on his own field, his first reaction is often one of dismay at the comparatively small amount of space that can be allotted to it. Writing articles for encyclopaedias is an art of its own; within a limited space so much must be compressed—nothing important can be omitted, nothing trivial should be included. Most experts would agree that it is easier to write a book than an encyclopaedia article. They would also consider it simpler to write an article for a periodical, because encyclopaedia writing is teamwork in which each article is edited in relation to others closely connected by subject. If a writer makes a statement that is partly qualified or totally contradicted in another article, the contributions of both writers must be scrutinized by the editorial staff, whose job it is to effect some kind of eventual agreement. Truth can be viewed from many standpoints, and references to any controversy may produce problems demanding all the skill and tact of the editors to resolve, particularly when the reputation of the writer is at stake in a signed article.

Revision and updating. The revision and updating of an encyclopaedia is one of the greatest challenges to its makers, and one to which many ingenious, if admittedly partial, solutions have been found. Louis Moréri set the example in his rapid incorporation of new information in each succeeding issue of his widely used *Grand Dictionnaire historique* ("The Great Historical Dictionary"; 1674). When the German publisher Friedrich Arnold Brockhaus first issued his great encyclopaedia he was forced by an unexpectedly large public demand to issue edition after edition in quick succession (some of them even overlapped). In all of these he took great pride in providing the latest information, personally supervising much of the revision of individual articles. Moreover, he provided special supplements incorporating these revisions for purchasers of each edition, so that they did not have to buy the next in order to keep up-to-date.

In the 18th and 19th centuries, most encyclopaedias that lasted long enough to require revision met the problem by preparing a new edition or by issuing supplements. In the case of *Encyclopædia Britannica,* the first edition (1768–71) was replaced by an essentially new and enlarged second edition in 1777–84; the ninth edition (1875–89), however, remained in print until the preparation of the 11th edition (1910–11), with a 10th edition nominally created by the addition of 11 supplementary volumes in the interim. Among the most serious shortcomings of the new-edition method was the tendency of publishers to dismiss editorial staff after the preparation of a new edition, a practice which meant that skilled editors were dispersed and had to be replaced once the decision to create a new edition had been taken.

Early in the 20th century it became the practice to fill the gaps between new editions with annual summaries called yearbooks (considered above). A turning point came when, soon after the publication of its 14th edition in 1929, *Encyclopædia Britannica* announced the introduction of a system of continuous revision that in one form or another is now the practice of most major encyclopaedias in many countries. Under continuous revision programs, some percentage of the articles in a set are updated or improved in other ways on a flexible schedule. Several publishers were able to take advantage of 20th-century printing technologies to reprint their sets on an annual basis and to introduce into each new printing as many revised entries as possible. The system implies the existence of a permanent editorial department able, with the assistance of academic advisers and article authors, to monitor the condition of entries on a constant basis.

Continuous revision

Although some publishers have reported annual revisions of 10 percent or more of the entries in their set, actual practice has varied so widely that averages have little meaning.

Nor is continuous revision without drawbacks. The most serious disadvantage may relate to the rapidity with which articles in a set become noticeably unbalanced in relation to one another. Changes and events requiring revision of articles are more readily apparent in the scientific, technological, biographical, and historical areas, with the result that articles in such fields are revised much more frequently than articles in such fields as the humanities, where important changes do occur though more subtly.

An equally important disadvantage in continuous revision has to do with the inherent difficulty of revising, on an article-by-article basis, a set of reference books containing many thousands of articles. In the first place, editors are usually unable to revise all the articles that might be affected by a new development. In the case of the assassination of a president, for instance, the editors of the next printing might add the event to his biography and even to the history of his country but be unable to acknowledge the event in all the other articles in which the president's name appears. In the second place, it is a fact that updating a single article is not always as simple as it might at first appear to be. In a biography, for instance, critical events can occur so often that it soon becomes no longer possible simply to add an additional sentence to the end of the piece: the death of the subject of the biography might

be the occasion for a reassessment of his significance or for the disclosure of long unknown or unpublicized information about him; in archaeology, a new discovery may be at serious variance with several previously held theories on which a whole article might well be based. In such instances, revision must go beyond the simple addition of a sentence or the insertion of a word or date and may involve partial or complete rewriting. With the pace of modern research what it is, this can quickly become an ever-present editorial problem of great complexity.

It will be seen that the problem of keeping an encyclopaedia up-to-date has two facets: the first, that of finding ways to assure that any one printing or edition is as up-to-date as possible at the time of its preparation, and, the second, that of finding ways to make it possible for purchasers to maintain the set in an up-to-date condition. Continuous revision is one example of an effort to come to grips with the first aspect of the problem; yearbooks represent another type of effort. One apparent answer to both aspects, the loose-leaf format, has never been a publishing success. Nelson's *Perpetual Loose Leaf Encyclopaedia* (second edition, 1920) was discontinued; the prestigious *Encyclopédie française* (1935–66), however, continues to be available in both loose-leaf and bound volumes.

Given the rapidity of change in the contemporary world, it was to be expected that encyclopaedia publishers would quickly seek ways of exploiting new technologies in the field of information storage, retrieval, and distribution in solving their revision problems. Though most early studies centred on the computer and related word-processing equipment, other technologies, including the so-called videodisk, have also been examined for possible applications. The immense storage capacity and ease of access to, and manipulation of, stored data have already made the computer an invaluable adjunct in the editorial departments of many encyclopaedia publishers. Several encyclopaedias are known to have been totally captured on tape for computer processing; in a few instances, the computer has become a major agent of editorial production by which article text is generated, changed as needed, and recaptured in its new forms.

Computer applications

In studying the applications of computer technologies, at least two U.S. publishers have been involved in experimental programs that make the entire texts of their sets available as an on-line computer service. Grolier, Inc., which in 1982 acquired *Academic American Encyclopedia* from Arete Publishing Co., has proposed to make its complete computerized text available to anyone subscribing to certain data-transmission services, such as *The New York Times'* Information Bank. In a separate development, Encyclopædia Britannica, Inc., made tapes of its 30-volume 15th edition and related yearbooks available for on-line use by subscribers to the services of Mead Data Central, a division of the Mead Corporation of New York. Early results of the experiments have not been announced. Should they succeed, the way will be opened to the development of an electronic encyclopaedia independent of alphabetical arrangement, accessible through methods other than the use of an index, and capable of being updated on an almost daily basis as new data are substituted for old.

Controversy and bias. Throughout the years most major encyclopaedias have been accused of reflecting bias in one or more of their articles. In the *Encyclopédie,* the lack of neutrality was intentional and apparent. Various editions of *Encyclopædia Britannica,* almost from the beginning, were accused of bias as well. The practice of relying on outside specialists for articles, a practice now followed by most serious encyclopaedias, has increased the likelihood that bias will be worked into an article. Many critics have felt that the reader is protected in such cases by the fact that the identity of the contributor is not hidden. It has also been argued that the presence of slanted opinions in an article gave to older encyclopaedias a colour and sense of conviction that is lacking in most modern works. Modern editors of major encyclopaedias, nevertheless, make every effort to eliminate any hint of bias in their sets, but the task is a difficult one. For example, an account of the Korean War would vary according to whether it was written by a North or South Korean, a Chinese, or an American writer. Similarly, the inclusion of a map showing the frontiers between two or more nations may give rise to vigorous controversy if the nations involved were to dispute any part of the boundaries as shown. The illustration of a painting with an attribution to one artist may draw strong protests from art critics who do not agree with the writer. Controversy today has grown rapidly on many subjects that were not in dispute a few years ago.

The kinds of encyclopaedias

GENERAL ENCYCLOPAEDIAS

It is now possible to see, in the last 2,000 years of encyclopaedia production, the existence of a pattern closely related to the changing social needs of each age. The outstanding circumstances that governed the policy and production of encyclopaedias for the first 15 centuries were that comparatively few people were able to read and, stemming partly from this and partly from the cost of materials and workmanship, that copies of any lengthy work were very expensive. Only when printing was introduced into Europe did the cost of production drop by any large amount; this development in turn helped to stimulate the growth of readership. A notable feature at the time of the early printing press was the sudden growth in the popularity of some of the older encyclopaedias as a result of the tendency to ensure a ready market by printing works of which many manuscript copies were in circulation.

Influence of printing

During the first 16 centuries of their publication the great majority of encyclopaedias comprised great anthologies of the most significant writings on as many subjects as possible. These excerpts were arranged in an order that was constantly varying according to the individual compiler's concept of the hierarchies of human knowledge; some of these classification systems were more suitable than others, but none was completely successful in meeting the tastes of the reading public because there was no general agreement on the essential order of ideas. Although the compiler exercised considerable latitude in his choice of items to include in his encyclopaedia, he often restricted comment to a minimum, so that the reader was free to form his own opinion of what was offered. In addition, because the compiler selected his material from what had already been written, the gaze of the reader was being turned to the past, and, although he could enjoy the heritage of the preceding cultures, he was not being put in touch with as much of the contemporary world as he might have desired.

Around the 10th or 11th century, a new type of encyclopaedia began to emerge, probably stimulated by the growing number of language dictionaries that, starting well before printing was used, grew ever more numerous once they could be produced by this method. Many early dictionaries were little more than enlarged glossaries, but from the time of *Suidas* onward, there began to appear a type of dictionary—now called encyclopaedic—that added to the definition and etymology of a word a description of the functions of the thing or idea it named. In some dictionaries, such as those of the Estiennes, a French family of bookdealers and printers, this description might in some cases be of considerable length. Thus the compilers of the new form of encyclopaedia that emerged in the 16th and 17th centuries inevitably thought in terms of arranging their entries in alphabetical order, because the dictionaries had already familiarized the reading public with this system. It needed only the success of a single encyclopaedia, such as Moréri's *Grand Dictionnaire historique,* to prove to publishers in general that this was a formula with considerable promise.

The last half of the 18th century brought such an upheaval in man's concept of the world that the time was ripe for further experiments in the form of the encyclopaedia. The French encyclopaedists Diderot and d'Alembert and their band of contributors broke no new ground in the physical format and arrangement of the encyclopaedia, but their work inspired the intelligentsia of other nations to produce really good encyclopaedias of their own. It is no coincidence that both the German *Brockhaus* and the Scottish *Britannica* appeared with policies so different from all that had gone before that no pub-

Formulas of *Brockhaus* and *Britannica*

lisher of any new encyclopaedia could afford to ignore their new patterns. Their formulas were so good that the modern encyclopaedia is simply a vastly improved elaboration of their method of arrangement and organization. The compilers of both encyclopaedias had taken the best ideas from the anthologies and miscellanies of the early period of encyclopaedia making and from the later stage of encyclopaedic dictionaries. Realizing that the reading public would not tolerate the omission of some subjects and the unequal treatment of others, they prepared works in which at least a few lines were devoted to almost every conceivable topic, and for more important subjects a full account was provided, written by an expert, if possible.

The three periods of the history of encyclopaedias—(1) to 1600, (2) 1601–1799, and (3) 1800 onward—are very unequal. They are, moreover, to a certain extent misleading, for the different forms of the encyclopaedia overlapped at each turning point for some years, and even today there are still some important survivals from the two earlier periods. One can study and compare what each of the three main types of encyclopaedia has had to offer by opening the *Encyclopédie française, Webster's Third New International Dictionary,* and the *Encyclopædia Britannica* at the places in which each describes the same subject. The *Encyclopédie française* will provide one or more well-written treatises on the subject by writers of note. This is exactly what the encyclopaedias of the earliest period offered; and in both the old and the contemporary encyclopaedia the reader is left free to form his own opinion after reading what the experts have to say. *Webster's,* a one-volume work, of course provides much less, but it also gives much more, because it adds definitions and, often, explanatory drawings or diagrams to an admirably concise text that tells the reader much in a very few lines. This is exactly what the encyclopaedic dictionaries of Louis Moréri, Antoine Furetière, and others were offering in the 17th and 18th centuries. The *Britannica*'s contribution is distinct from those of the other two in that it provides a synthesis of what is known on the subject to date, and attempts to assess its current position. The comparison of these three examples gives only an approximate idea of the nature of the encyclopaedia in each of its three main stages, but it also helps to remind the reader of today that ideas concerning the form of an encyclopaedia have evolved considerably throughout the centuries.

Background of readers before 1600

The encyclopaedias of the period before 1600 apparently were designed for a small group of people, who had much the same educational background as well as similar interests and opportunities to pursue them. In general, they had a common outlook on both religious and secular matters. Moreover, although they were citizens of many different countries, they were united by their knowledge and use of Latin, the international language.

The Eastern Roman emperor Constantine VII Porphyrogenitus (905–959) tried to plant firmly in the hearts of the most worthy of his contemporaries both knowledge and experience of the past. His were troubled times, and he felt justified in using much of his enforced leisure (he came to the throne in 911 but was not allowed to rule until 945) to provide for the administrators and emissaries of his court the most useful extracts from the writings of a very catholic selection of authors, including the patriarch of Constantinople John of Antioch (Johannes Scholasticus), the Roman historian Appian, the Greek historian Polybius, the Greek philosopher Socrates, the 5th-century Byzantine historian Zosimus, and many others. One of the unexpected by-products of this industry was the preservation of a large number of writings, a service that some of the other medieval encyclopaedias also performed.

An advantage of the encyclopaedists of the first period (*i.e.,* before 1600) was that each of them either knew or could visualize his reading public, a point that encouraged a minimum of commentary and moralizing. In a way, they were performing the duties of a personal librarian in that they drew their readers' attention to innumerable passages that they believed might be useful to them in their work or their private lives. The possibility of achieving even more was fully appreciated: the English scholar Alexander Neckham, in his early 13th-century *De naturis rerum* ("On the Natures of Things") hoped that by the imparting of knowledge he might help to lift or lighten man's spirit, and to this end he tried to maintain a simple and admirably clear text. Neckham's near-contemporary Bartholomaeus Anglicus similarly set himself in his *De proprietatibus rerum* ("On the Characteristics of Things") to bring to his readers' attention the nature and properties of the things and ideas on which the early Christian Fathers and the philosophers had expatiated, but he forbore to comment on their writings, leaving his readers to form their own judgments. The anonymous compiler of the *Compendium philosophiae* ("Compendium of Philosophy"; *c.* 1316) believed the knowledge of truth to be the supreme and final perfection of mankind; thus he never moralized on the contents of his encyclopaedia, its cumulative effect thereby being the more impressive.

Within the early period of the history of encyclopaedias a number of stages can be distinguished that make each group of works of great importance in any study of the development of scholarship and the spread of learning throughout the West. Encyclopaedias of classical times reached their culmination in Pliny's *Historia naturalis,* which was issued in the time of the Roman emperor Titus (AD 39–81). Not one of the encyclopaedias of Pliny or his predecessors paid much attention to religion; if it was discussed the approach was antiquarian, the gods of the different nations ruled by Rome being named and described in a dispassionate spirit that reflected both the tolerance and the noninvolvement of the Romans in these matters. The emphasis instead was on government, geography, zoology, medicine, history, and practical matters. The theories of the various philosophers were outlined impartially, no indication being given of any personal preference. This objective approach adopted by the Romans in their encyclopaedias was not achieved again until the 19th century.

Encyclopaedias of classical times

By the time of the Roman philosopher Boethius and the statesman Cassiodorus (*i.e.,* the 5th and 6th centuries AD), the position concerning objectivity had changed. Like Pliny and the Roman statesman Cato, Cassiodorus had been an administrator; and, while his predecessors had been engaged in interpreting and epitomizing the knowledge of the ancient world for the benefit of their own people, Cassiodorus realized the necessity for providing a new interpretation of this knowledge for the Goths, the new masters of Italy. In the next 700 years the impact of Christianity brought a new phase in Western encyclopaedia making, just as the impact of Islām is clearly visible in the Arabic encyclopaedias of the same period. Although religion is not always given pride of place in the encyclopaedias of those times, it pervades the whole of their contents. Thus Cassiodorus' division of his encyclopaedia into two main sections—divine and human—is made even more interesting by his inclusion of cosmography, the liberal arts, and medicine in the first section. Although the compilers of the encyclopaedias of this period could envisage in theory a perfectly logical arrangement for their encyclopaedias by starting with the creation and working downward to the smallest and least significant of God's creations, in practice they found this very difficult to apply, and the result was often only superficially scientific. Moreover, the inclusion of such topics as astrology and magic was surprisingly prevalent and only began to disappear after the publication of *Liber floridus* ("The Flowering Book"; *c.* 1120), by Lambert, a canon of Saint-Omer, a work that discarded practical matters in favour of metaphysical discussion.

Use of the vernacular

The third stage in the development of encyclopaedias came with the introduction of vernacular editions, such as the *Mappemonde* and *Li livres dou trésor,* and the reflection of the impact of Greek philosophical works (in translation) in the middle of the 13th century. In this era there was an increasing number of lay encyclopaedists—*e.g.,* Latini, Bandini, de la Torre—and the subject coverage changed to give more space and importance to the practical matters that interested the rising mercantile class. At the same time, theology no longer dominated the classification schemes. Humanism reached its full expression in the Spanish philosopher Juan Luis Vives' *De*

disciplinis (1531), in which all the compiler's arguments were grounded on nature and made no appeal to religious authority. The printing press had eliminated one of the most vexatious problems: the introduction or perpetuation of textual errors by the manuscript copyists. At the same time, the wider circulation of encyclopaedias through the unrestricted sales of printed copies brought about a situation in which the compilers could no longer envisage their reading public and accordingly adjusted their approach to their largely unknown audience.

ENCYCLOPAEDIC DICTIONARIES

The period spanning the 17th and 18th centuries is characterized by the flourishing of the encyclopaedic dictionaries that were pioneered by the Estienne family in France in the 16th century. During these two centuries this form of encyclopaedia reflected two different policies. On the one hand there was the encyclopaedia, such as those of the Germans Johann Theodor Jablonski and Johann Heinrich Zedler, that paid particular attention to the fields of history and biography. On the other, there was a new form of encyclopaedia—if the exception of the 12th-century *De diversis artibus* be set aside—that devoted itself to the arts and sciences. The first type can therefore be said to be retrospective in approach, while the arts and sciences encyclopaedia was clearly identifiable with contemporary matters.

None of these divisions is actually clear-cut, for many traditional encyclopaedias continued to be compiled throughout the period, and not all the historical-biographical encyclopaedias ignored the arts and sciences or contemporary people and events. Nevertheless, the issue of Antoine Furetière's encyclopaedia and the immediate follow-up by *Le Dictionnaire des arts et des sciences* (1694) by the writer Thomas Corneille (the younger brother of the playwright Pierre Corneille) were sufficient to indicate the growing public interest in a more modern form of encyclopaedia. This indication was confirmed by the successful publication of John Harris' *Lexicon Technicum* (1704), which the author described as "an universal English dictionary of arts and sciences: explaining not only the terms of art, but the arts themselves." It is significant that Harris omitted such subjects as theology, biography, and geography. The Englishman Ephraim Chambers went even further in describing his internationally influential *Cyclopaedia* (1728) as

> an universal dictionary of arts and sciences; containing an explication of the terms, and an account of the things signified thereby, in the several arts, both liberal and mechanical, and the several sciences, human and divine, compiled from the best authors.

Ideal contents as viewed in the 18th century

No century has seen more public discussion of the nature of the encyclopaedia than the 18th; at the same time, there was much uncertainty concerning its ideal contents. The fine Italian encyclopaedia of Gianfrancesco Pivati (the secretary of the Academy of Sciences at Venice), the *Nuovo dizionario scientifico e curioso, sacroprofano* ("New Scientific and Curious, Sacred-Profane Dictionary"; 1746–51), avoided the subject of history, whereas the German writer Philipp Balthasar Sinold von Schütz's *Reales Staats- und Zeitungs-Lexicon* ("Lexicon of Government and News"; 1704) concentrated on geography, theology, politics, and contemporary history and had to be supplemented by the German economist Paul Jacob Marperger's *Curieuses Natur-, Kunst-, Berg-, Gewerk- und Handlungslexikon* (1712; "Curious Natural, Artistic, Mining, Craft, and Commercial Encyclopaedia"), which covered the sciences, art, and commerce.

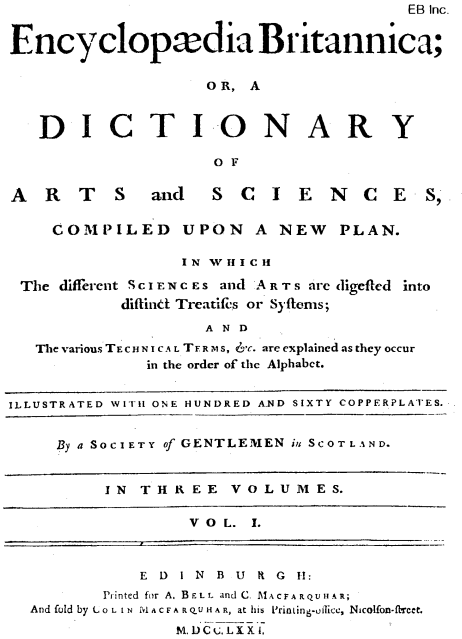
The introduction of the arts and sciences type of encyclopaedia inevitably hastened the use of specialist contributors, for it widened the total subject field considerably. "Hübner" (as Sinold von Schütz's encyclopaedia was known from the writer of the preface) employed many contributors, and it is known from the draft prospectus of the British writer Oliver Goldsmith that an encyclopaedia he projected was to have included comprehensive specialist articles by the lexicographer Samuel Johnson, the statesman Edmund Burke, the portrait painter Sir Joshua Reynolds, the historian Edward Gibbon, the economist Adam Smith, and others. The remarkable progress made in this period can easily be judged when one compares the encyclopaedia *Lucubrationes* (1541), in which the author, Joachim Sterck van Ringelbergh, found it necessary to include a "miscellaneous" section (which he amusingly dubbed "Chaos"), with the approach of Johann Georg Krünitz, a German physician and philosopher, in his highly organized, modern *Oekonomisch-technologische Encyklopädie* ("Economic-Technological Encyclopaedia"; 1773–1858) with its 242 volumes.

THE MODERN ENCYCLOPAEDIA

The period of the encyclopaedic dictionary was brilliant, but it gradually became apparent that, in abandoning the systematic encyclopaedia of the earlier period in favour of the quick reference dictionary form, quite as much had been lost as had been gained. The comparatively brief entries in the encyclopaedic dictionary had, by accident of the alphabet, fragmented knowledge to such an extent that users received only a disjointed knowledge of the things in which they were interested. Nor had the willful and extremely individualistic effort of the French encyclopaedists Diderot and d'Alembert done more than confuse the issue, for they had bent the principles of encyclopaedia making to their own purposes. An initial solution to the problem was found by Andrew Bell (1726–1809), Colin Macfarquhar (*c.* 1745–93), and William Smellie (1740–95), three Scotsmen who were responsible for the first edition (1768–71) of *Encyclopædia Britannica.* Aware of the shortcomings of the *Encyclopédie,* they devised a new plan. Their encyclopaedia was to include about 45 principal subjects (distinguished by titles printed across the whole page), supported by another 30 lengthy articles, the whole being contained within one alphabetical sequence interspersed with numerous brief entries enhanced by references, where appropriate, to the principal subjects. Some of the principal articles, notably those on medical subjects, extended to well over 100 pages each. The three collaborators had thus incorporated the comprehensive treatment of important subjects accorded by the earliest form of encyclopaedias and had supplemented this with the attraction of the brief informative notices of minor topics that had been the chief feature of the encyclopaedic dictionary. The key to their success was, however, their retention of the single alphabetical sequence.

Plan of the first *Encyclopædia Britannica*

Meanwhile, Renatus Gotthelf Löbel was planning to compile an encyclopaedia that could supersede "Hübner." It was Sinold von Schütz who, in the fourth edition


EB Inc.

Encyclopædia Britannica;

OR, A

DICTIONARY

OF

ARTS and SCIENCES,

COMPILED UPON A NEW PLAN.

IN WHICH

The different SCIENCES and ARTS are digested into distinct Treatises or Systems;

AND

The various TECHNICAL TERMS, &c. are explained as they occur in the order of the Alphabet.

ILLUSTRATED WITH ONE HUNDRED AND SIXTY COPPERPLATES.

By a SOCIETY *of* GENTLEMEN *in* SCOTLAND.

IN THREE VOLUMES.

VOL. I.

EDINBURGH:

Printed for A. BELL and C. MACFARQUHAR;

And sold by COLIN MACFARQUHAR, at his Printing-office, Nicolson-street.

M.DCC.LXXI.

Title page of volume 1 of the first edition of *Encyclopædia Britannica*, published in Edinburgh, 1768–71.

of "Hübner," had introduced the word *Conversations-Lexikon* into the title, and it was Löbel who decided to give it pride of place in his new encyclopaedia. The *Konversationslexikon* was designed to provide the rapidly growing German bourgeoisie with the background knowledge considered essential for entry into the polite society of the day. When Brockhaus took over Löbel's bankrupt and incomplete encyclopaedia, he saw the value and appeal of this evocative word and retained it (in various spellings) for many years afterward. Löbel's and Brockhaus' solution to the problem of the form of the modern encyclopaedia was not the same as the *Britannica*'s; it is interesting to note that whereas the *Britannica* model has widely prevailed throughout the English-speaking world, *Brockhaus* has been the model for most of the encyclopaedias prepared in countries in which English is not widely spoken.

Format of the *Konversationslexikon*

Brockhaus, throughout its existence, has faithfully followed a system in which the whole of knowledge has been analyzed into very specific topics. These topics are arranged alphabetically and, under each heading, condensed entries convey the essential information. By ingenious cross-references, entries are linked with other entries under which further information can be found, thus avoiding the inclusion of an index. There is no difficulty in distinguishing encyclopaedias of the *Konversationslexikon* form from encyclopaedic dictionaries. The former are usually of considerable size (*Der grosse Brockhaus,* 1928–35, included 200,000 articles by over 1,000 authors) and possess elaborate cross-reference schemes. Moreover, whenever a really important subject occurs, considerable space is allowed, though the same principle of concentrated text is followed.

Although the *Britannica* and *Brockhaus* examples eventually became the models for 19th- and 20th-century encyclopaedias, there have been many survivals from the previous periods. Ersch and Gruber's enormous *Allgemeine Encyclopädie* ("General Encyclopaedia"; 1818–89) has been cited as a true example of the medieval "summa"—it is famed for including the longest article in any encyclopaedia, that on Greece, which fills 3,668 pages in volumes 80–87. The *Encyclopédie française* is an even later example of this form and, as Samuel Taylor Coleridge planned it, the *Encyclopaedia Metropolitana* could have proved the supreme example of this type of treatment. Meanwhile, the encyclopaedic dictionary has never died, and at the very time when *Brockhaus* and the *Britannica* were building their markets, Noah Webster was developing his dictionary's reputation for reliability into a household expression in the United States.

ENCYCLOPAEDIAS FOR SPECIAL INTERESTS

Most encyclopaedias have been compiled from a purely scholarly point of view and have had no particular ax to grind, though nearly all have been inhibited to a certain extent by the interests and policies of the milieu in which they made their appearance. There are, however, several encyclopaedias that have deliberately been planned for a special purpose. One that is unique and continues to be of the greatest value to historians is the work of the 16th-century Spanish Franciscan Fray Bernardino de Sahagún, who spent much of his life in missionary work in Mexico. Sahagún was ordered to write in Nahuatl the information needed by his colleagues for the conversion of the Indians. The result, the *Historia general de las cosas de Nueva España* ("General History of the Matters of New Spain") is a magnificent record of the Aztec culture as recounted by the Indians of south central Mexico. The arrangement of this work, written in pictorial language as well as in Spanish, follows the familiar medieval pattern and resembles most closely that of Bartholomaeus Anglicus (Sahagún may have been familiar with a recent translation of Bartholomaeus' encyclopaedia). *Historia* is one of the most remarkable encyclopaedias ever compiled, and it has also a special importance for students of language.

Older Arabic and Chinese encyclopaedias

Many of both the Arabic and Chinese classical encyclopaedias were compiled with the object of helping civil service candidates in their studies and of providing administrators with the cultural background needed for their work. Their interest to historians of the two cultures can well be understood, for their arrangement and contents throw useful light on the concepts of administration and justice (to name only two aspects) in the Chinese and Islāmic worlds during the 7th to 15th centuries.

Of the Western medieval encyclopaedias, the most interesting in this respect is the *De naturis rerum* (*c.* 1228–44) of the Dominican friar Thomas de Cantimpré. His aim was that of St. Augustine: to unite in a single volume the whole of human knowledge concerning the nature of things, particularly the nature of animals, with a view toward utilizing it as an introduction to theology. To achieve this he devoted the first part of his work to mankind, zoology, botany, and geology. He then compiled a much smaller section on cosmology. An additional section on astronomy followed later. Although it was a comparatively unorganized and ill-balanced work, the popular nature of this encyclopaedia, with its stories of magic and allegorical creatures, ensured its widespread success over two centuries.

Religion and politics were the main motives for writing encyclopaedias with a special purpose. Louis Moréri made no secret of his intention to produce an encyclopaedia that would defend the teaching and policies of the Roman Catholic Church. Antoine Furetière and Pierre Bayle, on the other hand, represented the philosophers, and their anticlerical bias was more in tune with the skeptical minds of the age. Nevertheless, there was still a strong orthodox following in France, as the long-continuing demand for the *Dictionnaire universel* of the Jesuit fathers of Trévoux demonstrated, and this encyclopaedia was as firmly in defense of Catholicism as the *Encyclopédie* was critical of it.

Diderot and d'Alembert's encyclopaedia had originally been intended by its publisher to be no more than an adaptation of Ephraim Chambers' *Cyclopaedia.* The outcome was a giant reference work that criticized the government, satirized the Calvinist clergy of Geneva, championed the Age of Reason, and supported an atheistic materialism. To the more rigid members of the French Establishment, the encyclopaedia was a monster. But the more worldly had no objection to a work whose succeeding volumes were each an audacious source of scandal.

Objectivity in encyclopaedias

Even Pierre-Athanase Larousse, the French encyclopaedist, was not impartial. His finest encyclopaedia, the *Grand Dictionnaire universel du XIX^e siècle* ("Great Universal Dictionary of the 19th Century"; 1865–90), one of the most influential of the century, was deliberately anticlerical in policy. And Herder, in the heart of Catholic Germany, produced a counterweight to the Protestant *Brockhaus* in his *Konversations-Lexikon* (1853–57), which adopted a distinctive Catholic viewpoint. This excellent encyclopaedia was early recognized for its general impartiality, scholarship, and accuracy. In the long run, both "Herder" and *Brockhaus* gradually eliminated their sectarian inclinations, and the current editions are both of high standard.

CHILDREN'S ENCYCLOPAEDIAS

Before the 19th century, only Johann Wagenseil (1633–1705) had produced an encyclopaedia for children—the *Pera Librorum Juvenilium* ("Collection of Juvenile Books"; 1695). Larousse issued an interesting *Petite Encyclopédie du jeune âge* ("Small Children's Encyclopaedia") in 1853, but the next, *Encyclopédie Larousse des enfants* ("Larousse Encyclopaedia for Children"), did not appear until 1957. The first of the modern children's encyclopaedias was, however, a long-standing favourite. Prepared by the English writer and editor Arthur Mee (1875–1943), it was called *The Children's Encyclopaedia* (1910) in Great Britain and *The Book of Knowledge* (1912) in the U.S. The contents comprised a series of vividly written and profusely illustrated articles; because the system of article arrangement was obscure, much of the success of the work as a reference tool resulted from its splendidly contrived index, which even now remains a model of its kind. Arthur Mee later produced a completely pictorial encyclopaedia, *I See All* (1928–30), that comprised thousands of small illustrations, each accompanied by only a few words of text. Librarians still treasure it for its reference value, even if it is no longer used by children. In 1917–18 a completely new children's encyclopaedia was published,

The World Book Encyclopedia, which the title page described as "organized knowledge in story and picture." A success from the start, it issued enlarged editions in quick succession. In 1925 a volume devoted to reading courses and study units was added, and the index was abandoned in favour of innumerable cross-references in the text. Annual supplements were also provided from 1922 onward. In 1961 a Braille edition in 145 volumes was issued; most of the illustrations were eliminated in this, but many of the diagrams and graphs were retained and in many cases the captions for the pictures were incorporated in the text. In 1964 a separate 30-volume set in a special large type was published for the use of the partially blind.

World War I put a halt to the idea of issuing a *Britannica Junior,* and the first edition of such a work was not published until 1934. It was based on *Weedon's Modern Encyclopedia,* whose copyright had been bought by Britannica. Renamed *Britannica Junior Encyclopædia* in 1963, it was specifically designed for children in elementary-school grades. One of its features is its ready-reference index volume, which combines short fact entries with indexing to longer general articles. In 1960 a British *Children's Britannica* was issued in London. Prepared under the direction of John Armitage, London editor of *Encyclopædia Britannica,* its contents were determined largely by material covered in the so-called 11-plus standardized tests given in Britain. A yearbook supplement was added later, and the set has undergone periodic continuous revision since its first release.

In 1971 a new encyclopaedia, called *Young Children's Encyclopaedia,* was issued by Encyclopædia Britannica, Inc. Prepared specifically for children just learning to read and not yet into elementary school, it consists of 16 volumes, in which all the illustrations are in colour and the accompanying informative text is brief. Since its original appearance, the set has been translated into Japanese and Korean.

Compton's Encyclopedia

In 1894 Frank E. Compton sold a U.S. school encyclopaedia, the *Students Cyclopedia,* from door to door to pay his way through college. This later became the *New Students Reference Work,* which Compton finally bought. While continuing to publish this, Compton designed a completely new and, for those times, revolutionary work, which first appeared in 1922 as *Compton's Pictured Encyclopedia.* In due course, the system of continuous revision was introduced, close cooperation with educational and library advisers was fostered, and contributions from well-known authors were encouraged. In 1971 *Compton's,* by then published by Encyclopædia Britannica, Inc., introduced *Compton's Young Children's Precyclopedia,* based on the *Young Children's Encyclopaedia* described above.

Unlike *World Book, Compton's,* and the *Britannica Junior Encyclopædia,* the *Oxford Junior Encyclopaedia* (intended for children of 11 upward) is systematically arranged. Each of the 12 text volumes is devoted to a broad subject field: mankind, natural history, the universe, communications, great lives, farming and fisheries, industry and commerce, engineering, recreations, law and order, the home, and the arts. The 13th volume is an index, which includes a special section of ready-reference material. The contents of each volume are arranged alphabetically (with cross-references), and there are many illustrations.

SPECIALIZED ENCYCLOPAEDIAS

The alternative title of the 12th-century *Speculum universale* ("Universal Mirror") of a French preacher, Raoul Ardent (a follower of Gilbert de La Porrée, a French theologian), was the *Summa de vitiis et virtutibus* ("Summa [Exposition] of Faults and Virtues"). The title gives the clue of Raoul's intent, which was to provide a modern authoritative account of the Christian attitude to the world. His plan was different from that of other encyclopaedists, for he limited his work to the discussion (in this order) of theology, Christ and the redemption, the practical and ascetic life, thought, prayer, ethics, the four cardinal virtues, human conduct, and the four senses. This work could, in fact, be termed the first of the specialized encyclopaedias.

Apart from isolated examples such as this, and the technical encyclopaedia of Roger of Helmarshausen, the specialized encyclopaedia did not really make an appearance until the 18th century. The stimulus was probably provided by the increasing number of encyclopaedias that included the arts and sciences to such a point that some of them included little else. In any classified encyclopaedia the individual classes do, of course, constitute a kind of specialized encyclopaedia, but such a work falls far short of the real thing in that it is not sufficiently self-contained to stand on its own. It was inevitable, as the boundaries of knowledge contained in encyclopaedias grew greater, that there would be at least some attempts to produce specialized works of this kind; the French publishers Panckoucke and Agasse did something of this nature in their rearrangement of the contents of Diderot's *Encyclopédie* as a series of 196 volumes of separate dictionaries covering 41 subjects. The idea might have taken hold had it not been for insuperable difficulties caused by the French Revolution.

First specialized encyclopaedia on biography

The first real effort toward a specialized encyclopaedia had come some years earlier, and the subject field that it treated was biography. The *Allgemeines Gelehrten-Lexicon* ("General Scholarly Lexicon"; 1750–51) was compiled by Christian Gottlieb Jöcher, a German biographer, and issued by Gleditsch, the publisher of both "Hübner" and Marperger and the opponent of Zedler's encyclopaedia. Jöcher's work was continued by the German philologist Johann Cristoph Adelung and others and is still of value today. The field of international biography is not a simple one to tackle, and there were only two further efforts of note: J.C.F. Hoefer (1811–78) compiled the *Nouvelle Biographie générale* ("New General Biography"; 1852–66) and J.F. Michaud (1767–1839) was responsible for the *Biographie universelle.* These two great works were to a certain extent competitive; this helped to improve their coverage and content; they are still heavily used in all research libraries. After their publication, the task of recording biographical information on a universal scale reverted to the general encyclopaedias.

Specialized encyclopaedias of chemistry

Developments in the field of specialized encyclopaedias correspond closely to other developments in the world of scholarship. It is, for example, no accident that so much attention should be paid to the subject of chemistry at a time when L.F.F. von Crell was issuing his series of abstract journals on chemistry. The English scientist and inventor William Nicholson (1753–1815) was first in the field with his *Dictionary of Chemistry* (1795), published by Sir Richard Phillips (who later issued C.T. Watkin's *Portable Cyclopaedia*). On this was based Andrew Ure's *Dictionary of Chemistry,* which was for a long time the standard reference work on the subject in Great Britain. In 1807 the German chemist Martin Heinrich Klaproth issued his *Chemisches Wörterbuch* ("Chemical Dictionary"), but a more important event was the publication of the *Handbuch der theoretischen Chemie* ("Handbook of Theoretical Chemistry"; 1817–19) by the German scientist Leopold Gmelin, a work of such excellence that it still appears in new editions from the Gmelin-Institut. Heinrich Rose, a German chemist, issued his *Ausführliches Handbuch der analytischen Chemie* ("Complete Handbook of Analytic Chemistry") in 1851, and the first edition of the famous Liebig, Poggendorff, and Wöhler's *Handwörterbuch der reinen und angewandten Chemie* ("Handbook of Pure and Applied Chemistry") was issued in 1837; its second edition (1856–65) was expanded to nine volumes. This work was continued by Hermann Fehling's *Neues Handwörterbuch der Chemie* ("New Pocket Dictionary of Chemistry"; 1871–1930). The French counterpart, C.A. Wurtz's *Dictionnaire de chimie pure et appliquée* ("Dictionary of Pure and Applied Chemistry"; 1869–1908), became the standard work of its day. The Russian-born chemist Friedrich Konrad Beilstein first issued his *Handbuch der organischen Chemie* ("Handbook of Organic Chemistry") in Hamburg, Germany, in 1880–83; it is the most extensive work of its kind today. The French chemist Edmond Frémy's *Encyclopédie chimique* ("Chemical Encyclopaedia") appeared in 1882–99, and *A Dictionary of Applied Chemistry,* edited by Sir Thomas Edward Thorpe, the English chemist, was first issued in 1890–93. Twentieth-century standard works include Fritz Ullmann's *Enzyklopädie der technischen Chemie* ("Encyclopaedia of Applied

Specialized encyclopaedias of chemistry

Chemistry"; 1914–23), Victor Grignard's *Traité de chimie organique* ("Treatise on Organic Chemistry"; 1935), *Elsevier's Encyclopaedia of Organic Chemistry* (1940), the *Encyclopedia of Chemical Technology* (1947–56; known by the names of its principal editors as "Kirk-Othmer"), Waldemar Koglin's *Kurzes Handbuch der Chemie* ("Short Handbook of Chemistry"; 1951), and the indispensable *Handbook of Chemistry and Physics.*

The impressive run of encyclopaedias and handbooks of chemistry over so long a period is paralleled only in the field of music, in which the *Musikalisches Lexikon* ("Musical Lexicon"; 1732) of the German composer and music lexicographer Johann Gottfried Walther began the trend and was supplemented by the very successful *Historisch-biographisches Lexicon der Tonkünstler* ("Historical and Biographical Lexicon of Musicians"; 1790–92) of the German organist and music historian Ernst Ludwig Gerber. The *Biographie universelle des musiciens et bibliographie générale de la musique* ("Universal Biography of Musicians and General Bibliography of Music"; 1835–44) was compiled by the director of the Brussels Conservatoire, the Belgian composer François-Joseph Fétis, almost coinciding with the equally voluminous *Encyklopädie der gesammten musikalischen Wissenschaften* ("Encyclopaedia of Collected Musical Knowledge") of Gustav Schilling, a German lexicographer and historian of music. A pupil of Mendelssohn, Hermann Mendel, founded the *Musikalisches Conversations-Lexikon* (1870), which was completed by August Reissmann, who also edited the musicologist and composer Auguste Gathy's *Musikalisches Conversationslexikon* (1871). The great *Encyclopédie de la musique et dictionnaire du Conservatoire* (1913–31) was begun by the French writer on music Albert Lavignac and continued by Lionel de La Laurencie, but the third part, a dictionary of names and subjects covered in the preceding parts, was never issued. Walter Willson Cobbett compiled the *Cyclopedic Survey of Chamber Music* (1929–30), and the English writer on music Sir George Grove first issued his *Dictionary of Music and Musicians* in 1879–89; it went through five editions until a new work, *The New Grove Dictionary of Music and Musicians,* appeared in 1980. The German music historian Hugo Riemann compiled his standard *Musik-lexikon* in 1882, and the comprehensive *Musik in Geschichte und Gegenwart* ("Music of the Past and Present") began publication in 1949.

Encyclopaedias of philosophy

The publication of the German philosopher G.W.F. Hegel's *Encyklopädie der philosophischen Wissenschaften* ("Encyclopaedia of Philosophical Knowledge"; 1817) was of more than subject importance in that it was a compendium of the author's philosophical system in three parts; Logic, Nature, Mind. It influenced many editors of general encyclopaedias during the rest of the century. The standard work in this field has for many years been the *Dictionary of Philosophy and Psychology* edited by the American psychologist James Mark Baldwin, though the publication of *The Encyclopedia of Philosophy* (1967) provided a substantial work more in line with modern tastes. Other works in this area include the Centro di Studi Filosofici di Gallarate's *Enciclopedia filosofica* (1957), the French philosopher André Lalande's *Vocabulaire technique et critique de la philosophie* ("Technical and Critical Vocabulary of Philosophy", first issued 1902–12), and the Austrian writer Rudolph Eisler's *Wörterbuch der philosophischen Begriffe* ("Dictionary of Philosophical Concepts").

The Architectural Publication Society began issuing its *Dictionary of Architecture* as early as 1852, but it took 40 years to complete. A more modern work is *Wasmuths Lexikon der Baukunst* ("Wasmuth's Lexicon of Architecture"; 1929–37). Further material is included in the *Encyclopedia of World Art* (1959–68), the *Reallexikon für Antike und Christentum* ("Encyclopaedia for Antiquity and Christianity"; 1950–), and the *Enciclopedia dell'arte antica, classica e orientale* ("Encyclopaedia of Ancient, Classical, and Oriental Art"; 1958–66).

The words "Pauly-Wissowa" are very familiar to a great number of people. August von Pauly (1796–1845), the German classical philologist, began issuing his *Real-Encyclopädie der classischen Altertumswissenschaft* ("Encyclopaedia of Classical Antiquities") in 1837. The new edition was begun by another German classical philologist, Georg Wissowa, in 1893. This enormous work on classical studies has no equal in any part of the world, though it can be supplemented in some areas by the encyclopaedic series *Handbuch der Altertumswissenschaft* ("Handbook of Antiquities") begun in 1887.

The Swiss theologian J.J. Herzog (1805–82) gave religion its first great encyclopaedia with his *Real-Encyklopädie für protestantische Theologie und Kirche* ("Encyclopaedia of the Protestant Church and Theology"; 1854–68). Philip Schaff (1819–93), a Swiss-born U.S. church historian, prepared the abridged English edition (1882–84) from which *The New Schaff-Herzog Encyclopedia of Religious Knowledge* stems. James Hastings, a Scottish clergyman, was responsible for no fewer than four encyclopaedic works in this field: *A Dictionary of the Bible* (1898–1904); *A Dictionary of Christ and the Gospels* (1906–08); *Encyclopaedia of Religion and Ethics* (1908–26), still of great importance; and *Dictionary of the Apostolic Church* (1915–18). An even more significant series is the *Encyclopédie des sciences ecclésiastiques* ("Encyclopaedia of the Ecclesiastical Sciences"), which will take many decades to complete. It comprises: *Dictionnaire de la Bible* (1907–); *Dictionnaire de théologie catholique* (1909–); *Dictionnaire d'archéologie chrétienne et de liturgie* (1928–53); *Dictionnaire d'histoire et de géographie ecclésiastiques* (1929–); and *Dictionnaire de droit canonique* ("Dictionary of Canon Law"; 1935–). Other important works are *The Catholic Encyclopedia* (1907–18), which has not been completely superseded by the *New Catholic Encyclopedia* (1967); the finely illustrated *Enciclopedia cattolica* (1948–54); *Die Religion in Geschichte und Gegenwart* ("Religion in the Past and Present"; 1909–13); and the *Lexikon für Theologie und Kirche* ("Lexicon of Theology and the Church"; 1930–38). Two recently published encyclopaedias of religion include *The Encyclopaedia of Islam* (1960–) and the *Encyclopaedia Judaica* (1971–72).

Almanacs and handbooks of social and political science

It was not until the 1860s that three of the most useful handbooks now in daily use began to appear. *The Statesman's Year-Book,* important for its statistical and political information, began publication in 1864. In 1868 the English publisher Joseph Whitaker first issued his *Whitaker's Almanack,* and the *World Almanack* started in the same year. The *Chicago Daily News Almanac* appeared from 1885 to 1946, and the *Information Please Almanac* began in 1947. Herder's *Staatslexikon* ("Lexicon of Political Science") was first published in 1889–97; this compendium was soon followed by the *Dictionary of Political Economy* (1894) by the English banker and economist Sir Robert Palgrave. In 1930–35 the *Encyclopaedia of the Social Sciences* was published; an immediate success, it is often referred to as "Seligman" after the name of its chief editor. The new *International Encyclopedia of the Social Sciences* (1968) did not supersede it in every respect. In a similar fashion, the *Handwörterbuch der Sozialwissenschaften* ("Pocket Dictionary of the Social Sciences"; 1952–) supplemented rather than superseded the standard *Handwörterbuch der Staatswissenschaften* ("Pocket Dictionary of Political Science," 4th ed.; 1923–39).

In the field of literature, if Isaac Disraeli's *Curiosities of Literature* (1791) is ruled out, the first important handbook is the *Dictionary of Phrase and Fable* (1870) by the English clergyman and schoolmaster Ebenezer Cobham Brewer (1810–97), supplemented with Brewer's *Reader's Handbook* (1879). Other important works include the *Dizionario letterario Bompiani degli autori* ("Bompiani's Literary Dictionary of Authors"; 1956–57), the *Dizionario letterario Bompiani delle opere* ("Bompiani's Literary Dictionary of Works"; 1947–50), *Cassell's Encyclopaedia of Literature* (1953), and the Oxford "companions" to American, English, and French literature. In the last quarter of the 19th century, three major specialized encyclopaedias were issued: *Dictionnaire de botanique* ("Dictionary of Botany"; 1876–92) of the French naturalist and physician Henri Baillon, the *Lexikon der gesamten Technik* ("Lexicon of Collected Technology"; 1894–99) of the German engineer Otto Lueger, and the Berlin Academy's *Enzyklopädie der mathematische Wissenschaften* ("Encyclo-

paedia of Mathematical Sciences"; 1898–1935). The last was shortly followed by the important but incomplete *Encyclopédie des sciences mathématiques pures et appliquées* ("Encyclopaedia of Theoretical and Applied Mathematical Sciences"; 1904–14).

Physics never received the degree of attention that the encyclopaedists accorded to chemistry and chemical engineering. The standard *Dictionary of Applied Physics* of the English physicist Sir Richard Glazebrook was first issued 1922–23. The *Handbuch der Physik* ("Handbook of Physics") was also issued during those years; the second edition (1955) is often referred to by the name of its editor, Siegfried Flügge. The most recent work is the *Encyclopaedic Dictionary of Physics* (1961–64), edited by James Thewlis.

In medicine the pioneer *British Encyclopaedia of Medical Practice* (1936–39) has been followed by *The Encyclopaedia of General Practice* (1963).

Other important encyclopaedias and handbooks issued in recent years include *The Encyclopedia of Photography* (1949), the superbly illustrated and well-documented *Enciclopedia dello spettacolo* ("Encyclopaedia of the Stage"; 1954–62), which includes all forms of staged entertainment; the *Dictionnaire du cinéma et de la télévision* ("Dictionary of the Cinema and Television"; 1965–); the *McGraw-Hill Encyclopedia of Science and Technology* (1960), and the *Encyclopedia of Library and Information Science* (1968–).

ENCYCLOPAEDIAS OF COUNTRIES AND REGIONS

A special kind of encyclopaedia dealing with a single country or region began to appear in the late 19th century. Sometimes it is possible to distinguish, by a subtle form of titling, those national encyclopaedias that deal with the world scene from those that concentrate chiefly on their own country. Thus the "Ruritanian Encyclopaedia" can usually be taken to be a work produced in Ruritania that takes a world view, while the "Encyclopaedia of Ruritania" probably deals mainly with Ruritania and the surrounding areas. There are no rules in this matter, and this method of differentiating between the two kinds can only be regarded as an imprecise way of identifying the kind of encyclopaedia in the first instance.

The encyclopaedias of geography are of particular use in this field because they cover in detail many islands, small cities, and other features that are only dealt with in the briefest fashion elsewhere. Of the modern geographical encyclopaedias the following are of especial importance: *Westermanns Lexikon der Geographie* (1968–); *Meyers Kontinente und Meere* ("Meyer's Continents and Seas"; 1968–); the Russian *Kratkaya geograficheskaya entsiklopedya* ("Short Geographical Encyclopaedia"; 1960–66); and the *Länderlexikon* ("Geographical Dictionary"; 1953–60), of which a new edition began to be issued in the early 1970s. These encyclopaedias have an additional value as sources of maps and illustrations that would be difficult to find elsewhere. The main modern encyclopaedias dealing with continents, regions, and countries are:

Encyclopaedias of continents, regions, and countries

Africa: *Encyclopaedia of Southern Africa* (1961); *Deutsches Kolonial-Lexikon* (1920); *Encyclopédie coloniale et maritime* (1941–); *Grande encyclopédie de la Belgique et du Congo* (1938–52).

Australasia: *The Australian Encyclopaedia* (1958); *Encyclopaedia of Australia* (1968); *The Modern Encyclopaedia of Australia and New Zealand* (1964); *An Encyclopaedia of New Zealand* (1966).

The Americas: *Diccionario enciclopédico de las Américas* (1947); *Verbo: enciclopédia luso-brasileira de cultura* (1963–); *Diccionario enciclopédico del Peru* (1967); *Enciclopedia Yucatanense* (1944–47); *Enciclopedia de México* (1966–); *Diccionario histórico-enciclopédico de la Republica de El Salvador* (1927–); *Diccionario geográfico, estadístico, histórico, de la isla de Cuba* (1863–66); *Gran enciclopédia argentina* (1956–64); *Encyclopedie van de Nederlandse Antillen* (1969); *Encyclopaedie van Nederlandsch West-Indië* (1914–17); *Enciclopédia larense* (1941); *Encyclopedia Canadiana* (1957–58).

Europe: *Flandria nostra* (1957–60); *Magyar életrajzi lexikon* (1967–69); *Encyclopaedia of Ireland* (1968); *Latvijas PSR mazā enciklopēdija* (1967–); *Latvju enciklopēdija* (1950–55); *Mažoji lietuviškoji tarybinė enciklopedija* (1966); *Norge* (1963); *Enciclopedia românîei* (1938–43); *Encyclopédie polonaise* (1916–20); *Sverige: land och folk* (1966); *Ukraine: A Concise Encyclopaedia* (1963–); *Narodna enciklopedija srpsko-hrvatsko-slovenačka* (1925–29).

Middle East: *Eretz-Yisra'el: Entziqlopedia Ṭopografit-Hiṣṭorit* (1946–55).

History of encyclopaedias

ENCYCLOPAEDIAS IN THE WEST

Early development. The first fragments of an encyclopaedia to have survived are the work of Speusippus (died 339/338 BC), a nephew of Plato's. Speusippus conveyed his uncle's ideas in a series of writings on natural history, mathematics, philosophy, and so forth. Aristotle's wide-ranging lectures at the Lyceum were equally influential, and he and Plato appear to have been the originators of the encyclopaedia as a means of providing a comprehensive cultural background.

The Greek approach was to record the spoken word. The Romans, on the other hand, aimed to epitomize existing knowledge in readable form. Their first known effort is the *Praecepta ad filium* ("Advice to His Son"; *c.* 183 BC), a series of letters (now lost) written by the Roman consul Cato the Censor to his son. Cato's intention was to provide a summary of useful information that could help in the process of living and in guiding and helping one's fellowmen. A more substantial attempt was made by the learned Latin writer Marcus Terentius Varro in his *Disciplinarum libri IX* ("Nine Books of Disciplines"), his *Rerum divinarum et humanarum antiquitates* ("The Antiquities of Things Divine and Human"), and his *Imagines,* which together covered the liberal arts, human efforts, the gods, and biographies of the Greeks and Romans.

The most important Roman contribution was the *Historia naturalis* of Pliny the Elder, a vast work constituting a kind of classified anthology of information. Although undiscriminating in its record of fact and fancy, it was nevertheless very influential; the Latin grammarian and writer Gaius Julius Solinus drew nearly 90 percent of his 3rd-century *Collectanea rerum memorabilium* ("Collection of Memorabilia") from Pliny, and the *Historia naturalis* served as a major source for other encyclopaedias for at least the next 1,500 years. Even today it is still an important record for details of Roman sculpture and painting.

Pliny's *Historia naturalis*

The statesman Cassiodorus, when he withdrew to the Vivarium in 551, dedicated this monastery to sacred and classical learning. His *Institutiones divinarum et saecularium*

By courtesy of the trustees of the British Museum; photograph, R.B. Fleming & Co.

Illustration from the entry on the winds in St. Isidore of Seville's *Etymologiae,* an edition published in Strasbourg *c.* 1473.

litterarum ("Institutes of Divine and Secular Literature") seems to have been designed to preserve knowledge in times that were largely inimical to it. In his encyclopaedia, Cassiodorus drew a clear distinction between the sacred and the profane, but the first Christian encyclopaedia to be compiled for the benefit of the newly converted Spanish population followed a different scheme. St. Isidore (*c.* 560–636) considered the liberal arts and secular learning to be the true basis of a Christian's education. His *Etymologiae* therefore paid much attention to practical matters and even included an etymological dictionary. This was in line with the thought of St. Jerome—on whose encyclopaedic *Chronicon* and *De viris illustribus* St. Isidore had drawn—who, in common with the early Christian Fathers, was eager to provide a basis for a Christian interpretation and organization of knowledge. This concept was much later to be renewed by the Catalan ecclesiastic Ramon Llull.

The development of the encyclopaedia during the next 500 years, though of social interest, was undistinguished from the point of view of scholarship. Rabanus Maurus (*c.* 776–856), one of the English scholar Alcuin's favourite pupils, compiled *De universo* ("On the Universe"), which, in spite of its being an unintelligent plagiarism of St. Isidore's work, had a lasting popularity and influence throughout the medieval period. A series of encyclopaedias of special subjects—undistinguished anthologies of classical and Christian writings on history, jurisprudence, agriculture, medicine, veterinary surgery, and zoology—was organized by the Byzantine emperor Constantine VII Porphyrogenitus (905–959). Michael Psellus (1018–96), a tutor of a later emperor, contributed a more interesting work, *De omnifaria doctrina,* in the form of questions and answers on both the humanities and science. At this time there was a growing influence on metropolitan and secular learning. In an attempt to counterbalance it, the brief but charming *Didascalion* of Hugh of Saint-Victor (*c.* 1096–1141), which paid much attention to practical matters as well as to the liberal arts, was soundly based on a profound classification of knowledge that influenced many later encyclopaedias. About this time an encyclopaedic dictionary known as *Suda,* or *Suidas,* broke with tradition by adopting alphabetical order for its contents. This had no effect on the plan of later encyclopaedias, but its contents included so much useful information that it has retained its importance as a source throughout the succeeding centuries.

Adoption of the alphabetical order in *Suda*

The *Liber floridus* (*c.* 1120) of Lambert of Saint-Omer is an unoriginal miscellany, but it has an interest of its own in that it discards practical matters in favour of metaphysical discussion and pays special attention to such subjects as magic and astrology. The greatest achievement of the 12th century was the *Imago mundi* of Honorius Inclusus. Honorius produced his "mirror of the world" for Christian, later abbot of St. Jacob, and drew on a far wider range of authorities than any of his predecessors. The arrangement of the first section on geography, astrology, and astronomy was sound; it started with the creation and worked down to individual countries and cities. This was followed by a "chronicle," and a third section provided a brief list of important events since the fall of Satan. Honorius accurately foresaw his book's fate: innumerable copies, unauthorized plagiarisms, incessant criticism, and incompetent additions for at least 200 years.

Probably the first encyclopaedia to be compiled by a woman, the *Hortus deliciarum* of the abbess Herrad (died 1195), comprised a magnificent illuminated manuscript with 636 miniatures, intended to help and edify the nuns in her charge. Bartholomaeus Anglicus based his *De proprietatibus rerum* (1220–40) on the works of St. Isidore and Pliny. It was designed for ordinary people and became Europe's most popular encyclopaedia for the next three centuries. But the outstanding achievement of the Middle Ages was the *Speculum majus* of Vincent of Beauvais. Vincent was not an original writer but he was industrious, and his work comprised nearly 10,000 chapters in 80 books; no encyclopaedia rivalled it in size until the middle of the 18th century. The work was very well balanced, almost equal space being allotted to the three sections. The "Naturale" dealt with God and man, the creation, and natural history. For this Vincent drew not only on Latin writings but also on Greek, Arabic, and Hebrew, which were at that time (through translations) making a very considerable impact on the thinking of the West. The "Doctrinale" covered practical matters as well as the scholastic heritage of the age. The "Historiale" included a summary of the first two sections and a history of the world from the creation to the times of St. Louis. A fourth section, "Morale," based principally on St. Thomas Aquinas, was added after Vincent's death. The influence of the *Speculum majus* was immediate and lasting. Translations were made into several languages, and complete reprints appeared as late as 1863–79. One of its many values is that it is a source for extracts from many documents of which no other parts have survived. Another is its detailed history of the second quarter of the 13th century.

Vincent's was the last major work of its kind. Later encyclopaedists began to compile for a wider public than the very limited world of religious communities. The first breakaway from Latin came with *Li livres dou trésor* ("Treasure Books") of Brunetto Latini (*c.* 1220–95), the master of Dante, and the Florentine poet and philosopher Guido Cavalcanti. Latini wanted to reach the mercantile and cultured classes of Italy; he therefore used French, their common language. The arrangement of his work was similar to Vincent's but his approach was concise. The language, the brevity, and the accuracy of his encyclopaedia had an immediate and wide appeal. A friend of Petrarch's, Pierre Bersuire, based his *Reductorium, repertorium et dictionarium morale utriusque testamenti* ("Moral Abridgment, Catalogue and Dictionary of Each Testament"; *c.* 1340) on Bartholomaeus' *De proprietatibus rerum.* In contrast to Latini's work, this was a return to the traditional, with its moralizings on the Bible, Ovid's *Metamorphoses,* and natural history, but it had a considerable success when printing was introduced, being issued 12 times by 1526.

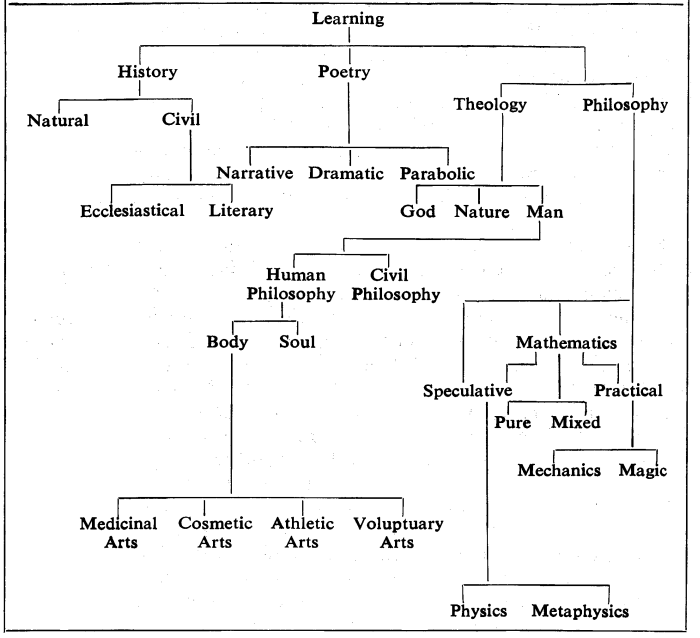
Brunetto Latini's breakaway from Latin

One of the most delightful of all encyclopaedias is the little *Margarita philosophica* that Gregor Reisch (died 1525) wrote for young people. In some 200 pages he contrived to cover in a very pleasing style the whole university course of the day, both the trivium and the quadrivium. The arrival of humanism is reflected in the *De disciplinis* of Juan Luis Vives, a pioneer in psychology and philosophical method; Vives grounded all his arguments on nature and made no appeal to religious authority. With the writing of the anonymous *Compendium philosophiae* (*c.* 1300) the concept of the modern scientific encyclopaedia was reached at last. It was the first encyclopaedia to adopt an inquiring and impartial attitude to the things described, and the old wives' tales that had filled so many pages of encyclopaedias from the time of Pliny onward were replaced by the latest scientific discoveries.

The first indigenous French encyclopaedia, the popular *Dictionarium historicum, geographicum et poeticum* ("Historical, Geographical, and Poetic Dictionary") of Charles Estienne (1504–64) was not published until 1553. For encyclopaedias in their own language, the French still had to rely on translations of the encyclopaedias of other nations, such as *Les diverses leçons* ("The Various Lessons"; 1552) of Pedro Mexia, a mediocre Spanish historian whose haphazard compilation was enormously popular in the 16th and 17th centuries.

The development of the modern encyclopaedia (17th–18th centuries). Francis Bacon's purpose in writing the *Instauratio magna* was "to commence a total reconstruction of sciences, arts, and all human knowledge, raised upon the proper foundations" in order to restore or cultivate a just and legitimate familiarity between things and the mind. Only a small part of this enormous work was ever completed, but the author had planned 130 sections divided into three main sections: external nature, man, and man's action on nature. From its proposed contents Bacon's intention was clearly to compile an encyclopaedia thoroughly scientific in character—"a thing infinite and beyond the powers of man"—that he himself recognized to be revolutionary in character. His most important contribution was, however, the devising of a new and thoroughly

Table 1: Francis Bacon's Classification of Knowledge



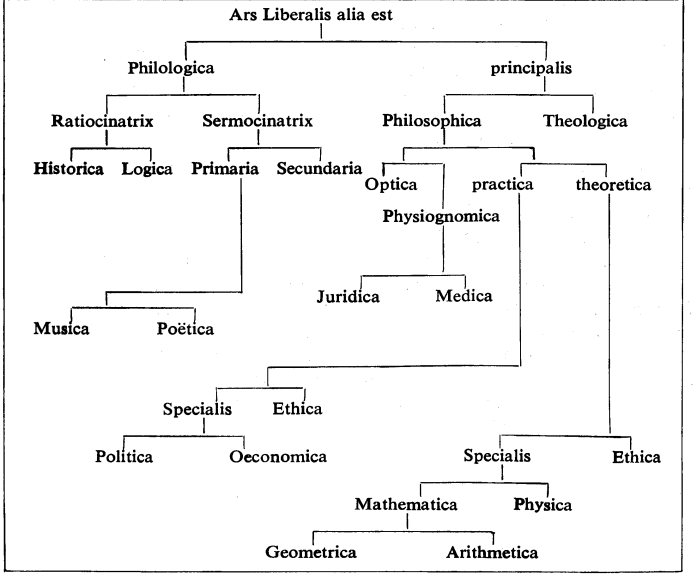
sound classification of knowledge that bears a remarkable resemblance to the classification put forward by Matthias Martini in his *Idea Methodica* (1606). Although Bacon was apparently unaware of this work, both philosophers were probably working from the same basic Platonic precepts. The results were profound: Diderot made a point of acknowledging the assistance Bacon's analysis of the structure of human knowledge had afforded him in planning the contents of the *Encyclopédie,* and Samuel Taylor Coleridge hailed "the coinciding precepts of the Athenian Verulam and the British Plato."

Bacon's influential classification

Only two more Latin encyclopaedias of any importance followed. Antonio Zara, bishop of Petina, compiled the *Anatomia Ingeniorum et Scientiarum* ("Anatomy of Arts and Sciences"; 1614), which was chiefly remarkable for the inclusion of an index. And Johann Heinrich Alsted, who, like Martini, came from Herborn, compiled an *Encyclopaedia* (1630) whose arrangement corresponds broadly to Matthias' classification of human knowledge.

Zara's and Alsted's encyclopaedias were organized systematically by classification. The turning point came with

Table 2: Matthias Martini's Classification of Human Knowledge



Louis Moréri's alphabetically arranged *Grand Dictionnaire historique* (1674), which was especially strong in geographical and biographical material. Its success was immediate; six editions were issued by 1691, each incorporating much new contemporary information. English editions followed in 1694, 1701, and (a supplement) 1705. Other encyclopaedias in England, Germany, Switzerland, and The Netherlands acknowledged its inspiration. The alphabetically arranged encyclopaedia in the vernacular had almost won the day, in spite of the German scholar Daniel George Morhof's modest success with his ill-balanced *Polyhistor Literarius, Philosophicus, et Practicus* ("Literary, Philosophical, and Practical History"; 1688–1708).

If there was any doubt concerning the more popular form of the encyclopaedia, the issue of Antoine Furetière's *Dictionnaire universel des arts et sciences* (1690) confirmed the true nature of public taste. Furetière not only compiled a fine encyclopaedic dictionary, but he emphasized the arts and the sciences, thus reflecting the rapidly growing public interest in modern culture, science, and technology. If confirmation were still needed, the Académie Française's commissioning of Thomas Corneille to compile *Le Dictionnaire des arts et des sciences* (1694), with its thorough and authoritative treatment of these new encyclopaedic features, demonstrated that even the more conservative scholars were by now keenly aware that a new spirit had arisen. The period of the clerical encyclopaedia had ended, as the Franciscan friar Vincenzo Maria Coronelli found when his *Biblioteca Universale Sacro-Profano* (1701–06) ceased publication at volume 7 of a projected 45.

Furetière's modern *Dictionnaire universel*

Pierre Bayle, in his *Dictionnaire historique et critique* (1697), achieved a most remarkable tour de force. Although his encyclopaedia purported to be an updating of the information in Moréri, the entries were largely unexceptionable. The real originality of his work lies in the profuse and scholarly footnotes and the commentaries that at times were an amazing mixture of skepticism, blasphemy, and ribaldry. Bayle challenged orthodox ideas; his brilliant mind spared nothing. This approach heralded that of Diderot, and the distinguished writers who revised later editions—Prosper Marchand and Pierre Desmaizeaux—continued in the same style.

The *Lexicon Technicum* (1704) of John Harris represented the powerful impact of the work of the Royal Society (founded 1660). Here was all the equipment of the modern encyclopaedia: excellent engraved plates, clear practical text, bibliographies appended to the more important articles. So far, England had had to make do with translations of French encyclopaedias. Harris' emphasis on the need to include scientific and technical subjects helped to reverse the trend. This process was completed by the issue of Ephraim Chambers' *Cyclopaedia* (1728). Like Harris, Chambers omitted people in favour of more information on the arts and sciences, and he paid more attention to clear expositions of ancient and modern philosophical systems. His admirably cross-referenced work is universally recognized as the father of the modern encyclopaedia.

The French were well aware of these developments. By 1744 five editions of Chambers' *Cyclopaedia* had been issued. The Paris publisher André Le Breton saw a ready market for a translation. The first proposals were a failure, however, and Diderot was enlisted to plan what at that time was still essentially a translation on a much broader basis. Under the hands of Diderot and d'Alembert the concept changed. The *Encyclopédie* (1751–65) was a philosophical undertaking carried out on a gigantic scale, and much of the writing was of a high standard. To the orthodox, it appeared that the project had got out of hand, but there were 2,000 subscribers to the first volume, and the subsequent scandals over the irreverent, authority-challenging articles only added to the number of purchasers. The equivocal attitude of high dignitaries in both church and court and the growing public dislike of the encyclopaedia's chief critics—the Jesuits—led to a complex situation in which official disapproval and substantial private encouragement caused the production and fortunes of the *Encyclopédie* and its producers to lurch dangerously from one crisis to another. Curiously, Diderot

Diderot's monumental *Encyclopédie*

did nothing to further the physical development of the encyclopaedia; his contribution was to fire men's minds with a willful guidance that conformed to the country's increasingly revolutionary spirit. As Voltaire said: "this vast and immortal work seems to reproach mankind's brief life span."

The shortcomings of the *Encyclopédie* were obvious. The essential ingredients of an encyclopaedia, the entries on every conceivable subject, had been sacrificed to make place for lengthy polemics on the controversial topics of the day. The *Encyclopædia Britannica* was intended to improve on this, and, with all its shortcomings, the first edition (1768–71) did exactly that. The achievement of its editors was the more remarkable in that there were already several English encyclopaedias on the market. The Scottish encyclopaedia, however, reflected the taste of the day better than any of its competitors, for it was a completely new work and not just a remaking of Chambers and Harris. There was much to criticize in the first edition, but the second (1777–84; dated 1778–83) was greatly improved, and succeeding issues of the *Britannica* made steady progress.

Meanwhile, Germany, at first largely dependent on translations of foreign encyclopaedias, had produced the scholarly "Hübner" (1704), as it was known from the name of the author of the preface in this first of the *Konversationslexikon* type. The form appealed to the rapidly growing middle class of the country, who welcomed encyclopaedias designed to provide them with an adequate cultural background for polite society. Johann Theodor Jablonski's illustrated *Allgemeines Lexicon* (1721) continued in this same style, and there were similar works compiled by the Swiss theologian and philologist Jakob Christoph Iselin and Antonius Moratori (1727). Johann Heinrich Zedler's huge *Grosses vollständiges Universal-Lexicon* ("The Great Comprehensive Universal Lexicon"; 1732–50) was in the older tradition but is important for its accuracy and its biographical and bibliographical material. An attempt to produce a German type of the *Encyclopédie* in 1778–1807 was, however, a failure. Friedrich Arnold Brockhaus recognized the real need of the German people. Reworking Renatus Gotthelf Löbel's bankrupt encyclopaedia, he produced his first *Konversations-Lexikon* (1796–1811), thereby setting the pattern for at least half of all succeeding encyclopaedias throughout the western world. Brief, well-designed articles tightly packed with facts, comprehensive coverage, and a reputation for accuracy and up-to-dateness were the ingredients for one of the most successful of encyclopaedias.

Brockhaus' Konversations-Lexikon

The 19th century. Having served a long apprenticeship as a reviser of Chambers' *Cyclopaedia,* Abraham Rees at last produced a completely original and finely illustrated work, *The New Cyclopaedia* (1802–20), the only serious rival to the *Britannica* in a generation that saw some dozen "new" encyclopaedias rise and fall. What might have been the greatest encyclopaedia of the century, the *Encyclopaedia Metropolitana* (1817–45), failed miserably because of the early withdrawal of its designer, Samuel Taylor Coleridge, and subsequent financial troubles; but from it came the most notable contribution to the philosophy of encyclopaedia making since Bacon—Coleridge's profound treatise "On Method" (1818).

To the principal influences on the compilation of encyclopaedias—Bacon, Diderot, the *Britannica,* and Brockhaus—must be added that of the Frenchman Pierre Larousse. His completely original approach to encyclopaedia making has given the series of encyclopaedias that bear his name a unique reputation. Emphasis throughout has been on readability; style has never been sacrificed to conciseness, and the successive editors of *Larousse* have paid very close attention to the changing public taste among French readers concerning the presentation of information.

The advent of Noah Webster was fully as epoch-making as that of Brockhaus and Larousse. Webster's informative *American Dictionary of the English Language* (1828) was encyclopaedic in character, but he avoided the long entries for the more important subjects that were such a feature of *Larousse.* Webster's approach appealed to the American taste and captured a huge market that has only increased with the years.

Brockhaus soon faced opposition, for his encyclopaedia was stronger on the humanities than on scientific and technical subjects. Joseph Meyer's *Der grosse Conversations-Lexikon* (1840–52) rectified this imbalance and was the first of a highly successful series that competed vigorously with *Brockhaus* for 100 years. In addition, Herder's *Conversations-Lexikon* (1853–57) and its subsequent editions provided the Catholic counterbalance in a country where Protestants and Catholics were almost equal in numbers.

The market for encyclopaedias in 19th-century Great Britain seemed inexhaustible, but many publishers lost money by putting out works that failed to capture the public's fancy. An exception was *Chambers's Encyclopaedia* (1860–68), which was unconnected with Ephraim Chambers' classic. Influenced by childhood access to a copy of the *Britannica,* Robert Chambers and his brother William compiled an original work, *Chambers's Encyclopaedia,* that took the *Konversationslexikon* form and thus found a new market that has continued to the present day.

Chambers's Encyclopaedia in England

In the first half of the 19th century there was increasing activity in other countries. Poland produced the *Encyklopedia Powszechna* (1858–68), known as "Orgelbrand" after its publisher. The Hungarians had followed the Bohemian *Slovník naučný* ("Scientific Dictionary"; 1860–90) with the *Egyetemes magyar encyclopaedia* ("Universal Hungarian Encyclopaedia"; 1861–76). The Russians had produced half an encyclopaedia, V.N. Tatishchev's *Leksikon rossyskoy* ("Russian lexicon"), in 1793, and then issued A. Starchevsky's *Spravochny entsiklopedichesky slovar* ("Encyclopaedic Reference Dictionary"; 1847–55) on the *Brockhaus* model. More important was the famous *Entsiklopedichesky slovar* ("Encyclopaedic Dictionary"; 1895), which became known as "Granat" after the Granat Russian Bibliographical Institute that produced it. A later edition (1910–48) of "Granat," in 58 volumes, is not exported from the Soviet Union. Modelled on the *Britannica,* this edition contained many important articles, such as Lenin's contribution on "Marx" and on "The Russian 19th-Century Agrarian Problem." Successive ideological changes in Russian society caused many changes in the text of "Granat," and it remains one of the most inaccessible of all Russian encyclopaedias outside the Soviet Union.

Larousse did not go unchallenged. Inspired by the French politician Ferdinand-Camille Dreyfus, *La Grande Encyclopédie* (1886–1902) provided France with a superb, authoritative, and comprehensive work, well documented, and of excellent scholarship throughout; it is still in use today. In Denmark the century ended with the issue of no fewer than three new good multivolume encyclopaedias: *Allers* (1892–99), *Hagerups* (1892–1900), and *Salmonsens* (1893–1911), a situation without parallel in the history of encyclopaedias. During the course of the century practically every feature of the modern encyclopaedia had been introduced, and editorial standards had at times risen to a height that modern editors can only envy.

The 20th century. In 1890–1906 a Russian edition of *Brockhaus,* which subsequently had considerable success, was issued from the St. Petersburg office of *Brockhaus.* In contrast, S.N. Yushakov designed his *Bolshaya entsiklopedya* ("Great Encyclopaedia"; 1900–09) on the "Meyer" model. After "Granat" the next important encyclopaedia was the 65-volume *Bolshaya sovetskaya entsiklopedya* ("Great Soviet Encyclopaedia"; 1926–47), which was eventually discredited; the second edition (1949–58) had a Marxist-Leninist approach but was less biassed on nonpolitical subjects. It represented almost the whole of the Soviet Union's cultural resources: 8,000 scholars contributed articles, and the appended bibliographies were truly international in scope. One complete volume was devoted to the Soviet Union. The yearbooks that supplemented this encyclopaedia were very well produced and maintained the high standards of the original work. From 1970 to 1978 a 30-volume third edition was issued. The reduction in size was accomplished by editing and the use of a smaller typeface. Early reviews indicated that the quality of the work was similar to that of the second edition.

Russian encyclopaedias

From 1973 to 1982 Macmillan released an English translation of the set.

There has also been a series of editions of the much smaller *Malaya sovetskaya entsiklopedya* ("The Little Soviet Encyclopaedia"), first issued in 1928–31.

In the U.S., the first edition of the *New International Encyclopaedia* was issued in 1902–04 and was subsequently supplemented by yearbooks. *The Encyclopedia Americana,* which traces its ancestry back to an English-language adaptation (1829–33) of the seventh edition of *Brockhaus,* took on new strength in 1902 when the editor of *Scientific American,* Frederick C. Beach, was appointed editor of the *Americana.* It has enjoyed growing success through its policy of following the continuous revision system, and yearbooks have supplemented it from 1923 onward. In 1950–51 a completely new American work, *Collier's Encyclopedia,* appeared in 20 volumes, and subsequent editions have been supplemented by yearbooks since 1960. *Collier's* is noted for its large number of illustrations and maps.

The "Espasa," the *Enciclopedia universal ilustrada europeo-americana* (1905–70), is a great encyclopaedia, which—like the *Enciclopedia italiana*—eschews revision in favour of a series of sizable supplements. Physically, it is the largest current encyclopaedia in the world. One complete volume is devoted to Spain and is separately revised and reissued from time to time. A smaller encyclopaedia, the *Diccionario Salvat* (first issued in 1907–13), is revised at frequent intervals. Another major Spanish encyclopaedia, the *Enciclopedia labor* (1955–60), is remarkable for devoting each volume to a major subject field, an index volume providing the key to the total contents. This encyclopaedia is valuable for the attention it pays to every part of the Spanish-speaking world.

Italian encyclopaedias

One of the most important of all encyclopaedias, the *Enciclopedia italiana* (1929–36), is famous for its lavish production, its superb illustrations, and its lengthy, scholarly, and well-documented articles. Even its defense of Fascist ideology is not allowed to impinge on the general impartiality of the text. Supplements have been issued since World War II. The postwar *Dizionario enciclopedico italiano* (1955–63), issued by the same publishers, is a much smaller but well-illustrated work. The *Enciclopedia Europe* was released in Milan between 1976 and 1981. Although consisting largely of brief articles, it has numerous signed long articles of good quality. In Germany, the three giants of the German encyclopaedia world—*Brockhaus,* "Meyer," "Herder"—continue to produce new editions, unchallenged by any new encyclopaedia of comparable scope.

In spite of the continuing popularity of *Larousse,* France has produced three other encyclopaedias of note in the 20th century. The *Encyclopédie française* (1935–66) is an outstanding collection of monographs by well-known scholars and specialists, arranged in classified form and available in loose-leaf binders, supplemented by a continuously revised index. Its 21 volumes, each under the direction of a different authority, deal with (1) man's mental tools (logical thought, language, and mathematics); (2) physics; (3) heaven and earth; (4) life; (5) living beings; (6) human beings (the normal man and the sick man); (7) the human species; (8) the study of the mind; (9) the economic and social universe; (10) the state; (11) international life; (12) chemical science and industry; (13) industry and agriculture; (14) man's daily life; (15) education and learning theory; (16, 17) arts and literatures, in two volumes; (18) the written word; (19) philosophy and religion; and (20) the world in its development (history, evolution, prospective); the 21st volume contains an index. The articles are notable for their almost total concentration on contemporary issues in the fields considered.

The *Encyclopédie de la Pléiade* (1955–) is an encyclopaedic series, each work (some in more than one volume) being a self-contained treatment of a broad subject field written in narrative form.

One of the most interesting new encyclopaedias is the *Encyclopaedia universalis* (1968–75), edited by Claude Grégory and owned by the French Book Club and Encyclopædia Britannica, Inc. This work, inspired by *L'Encyclopédie,* eschews the inclusion of minor items in favour of extensive and very well-illustrated articles on important subjects, and it pays special attention to modern science and technology. Of the 20 volumes (30,000,000 words), volumes 17–19 constitute a "thesaurus" of fact entries and index, and volume 20 is devoted to an exposition of the classifications used in constructing the encyclopaedia and to further elaborations. *Encyclopaedia universalis* is doubly notable as the product of a contemporary publishing phenomenon known in the industry as "coproduction." The term is applied in general to the collaborative efforts of publishers in two or more countries who have united to produce an encyclopaedia for sale in one of the countries or, with modifications in two or several. Successful examples of coproduction include the *Buritanika Kokusai Dai Hyakka Jiten* (or "Britannica International Encyclopædia") in Japan and the new Chinese encyclopaedia (both discussed elsewhere in this article). Encyclopædia Britannica, Inc., has, in addition, been similarly involved in the development of the Spanish-language *Enciclopedia Barsa* and the Portuguese-language *Enciclopédia Barsa,* each in 16 volumes; *Enciclopédia Mirador Internacional,* a scholarly 20-volume set first published in Brazil in 1975; and *Il Modulo,* a 24-volume set published in Italy. Other major instances of coproduction involve the interesting *The New Caxton Encyclopedia,* which originated in Italy with Instituto Geografico de Agostini and subsequently appeared in Great Britain, first in serial parts as *Purnell's New English Encyclopedia* (1965) and then in a bound set of 18 volumes (1966); in France there appeared a version called *Alpha: La Grande Encyclopédie Universelle en Couleurs,* and in Spain a version called *Monitor.* The American-made *The Random House Encyclopedia* has been adapted in various languages and under various names for distribution in several countries.

Co-production

Most countries have issued at least one good encyclopaedia, among which should be mentioned:

Bulgaria: *Kratka bŭlgarska entsiklopediia* (1963–).

Czechoslovakia: *Ottův slovník naučný* (1888–1909); *Masarykův slovník naučný* (1925–33); *Nový velký ilustrovaný slovník naučný* (1929–34); *Komenského slovník naučný* (1937–38); *Příruční slovník naučný* (1962–67).

Denmark: *Gyldendals Nye leksikon* (first issued 1931–32); *Raunkjaers konversations leksikon* (1948–57); *Gyldendals store opslagsbog* (1967–70).

Estonian S.S.R.: *Eesti nõukogude entsüklopeedia* (1968–).

Finland: *Otavan iso tietosanakirja/Encyclopaedia fennica* (1960–65); *Nuorten tieto* (1963–); *Fokus: Otavan kertovasti kuvitettu tietosanakirja* (1963–).

Greece: *Megalē hellenikē enkyklopaideia* (1926–34); *Eleutheroudakē enkyklopaidikon leksikon* (1927–31).

Hungary: *Révai nagy lexikona* (1911–35); *Új magyar lexikon* (1959–62); *Új idők lexikona* (1936–42).

The Netherlands: *Eerste nederlandse systematisch ingerichte encyclopaedie* ("ENSIE"; 1946–52); *Grote Winkler Prins encyclopedie* (1966– ; the latest in an outstanding series of "Winkler Prins" encyclopaedias starting in 1870–82); *Algemene nederlandse systematisch ingerichte encyclopedie* ("ANSIE"; 1955–); *Oosthoeks encyclopedie* (1968–); the latest edition of a series starting in 1916–23).

Norway: *Aschehougs Konversasjons leksikon* (1970– ; the latest in a series started in 1907–13); *Norsk konversasjonsleksikon* (first published 1931–34); *Norsk allkunnebok* (1948–61).

Poland: *Wielka Encyklopedia Powszechna PWN* (1962–).

Romania: the Romanian Academy's *Dictionar enciclopedic român* (1962–66).

Sweden: *Bonniers konversations lexikon* (first published 1922–32); *Focus uppslagsbok* (1958–60).

Turkey: *Türk ansiklopedisi* (1946– ; first called *Inönü ansiklopedisi*).

Yugoslavia: *Enciklopedija Jugoslavije* (1955–); *Enciklopedija leksikografskog zavoda* (first published 1955); *Pomorska Enciklopedija* (1954–64); *Hrvatska enciklopedija* (1941–45; unfinished, A–El only).

Lithuania (published in Boston): *Lietuviu enciklopedija* (1953–69).

ENCYCLOPAEDIAS IN THE EAST

China. The contribution from the East to the history of encyclopaedias is distinctive and covers a longer period than that of the West. The Chinese have produced encyclopaedias for approximately 2,000 years, but traditionally they differ from the modern Western encyclopaedia in that they are mainly anthologies of significant literature with some elements of the dictionary. Compiled by scholars of eminence, they have been revised rather than replaced over hundreds of years. In the main, they followed a classified form of arrangement; very often their chief use was to aid candidates for the civil service. The first known Chinese encyclopaedia, the *Huang-lan* ("Emperor's Mirror"), was prepared by order of the emperor in about AD 220. No part of this work has survived. Part of the *Pien-chu* ("Stringed Pearls of Literature"), prepared around 600, is still extant. About 620 the *I-wen lei-chü* ("Anthology of Art and Literature") was prepared by Ou-yang Hsün (557–641) in 100 chapters divided into 47 sections. The *Pei-t'ang shu-ch'ao* ("Extracts for Books") of Yü Shih-nan (558–638) was more substantial and paid particular attention to details of the organization of public administration. An annotated edition, edited by K'ung Kuang-t'ao, was published in 1880.

The earliest Chinese encyclopaedia

The *Ch'u-hsüeh chi* ("Entry into Learning") was a modest work compiled about 700 by Hsü Chien (659–729) and his colleagues. A more important book was the *T'ung-tien* ("Comprehensive Statutes") compiled by Tu Yu (735–812), a writer on government and economics. Completed in about 801, it contained nine sections: economics, examinations and degrees, government, rites and ceremonies, music, the army, law, political geography, national defense. In 1273 it was supplemented by Ma Tuan-lin's enormous and highly regarded *Wen hsien t'ung k'ao* ("General Study of the Literary Remains"), which included a good bibliography. Supplements to this work were published in the 17th, 18th, and 20th centuries. Under the order of the second Sung emperor, Sung T'ai Tsung, the statesman Li Fang organized the compilation of the vast *T'ai-p'ing yü-lan* ("Emperor's Mirror"), which included extracts from many works of literary and scientific standing that are no longer extant. In 1568–72 the *T'ai-p'ing yü-lan* was revised and reprinted from movable type; a new edition revised by Yüan Yüan appeared in 1812. The *Ts'e-fu yüan-kuei* (*c.* 1013), particularly strong in historical and biographical subjects, was almost as large as the *T'ai-p'ing yü-lan.*

The historian Cheng Ch'iao (1108–66) compiled the *T'ung chih* ("General Treatises"), an original work with a strong personal contribution; the printed edition (1747) was in 118 volumes. One of the richest and most important of all Chinese encyclopaedias, the *Yü-hai* ("Sea of Jade"), was compiled about 1267 by the renowned Sung scholar Wang Ying-lin (1223–92) and was reprinted in 240 volumes in 1738.

World's largest encyclopaedia

What was probably the largest encyclopaedia ever compiled, the *Yung-lo ta-tien* ("Great Handbook"), was issued at the beginning of the 15th century. Unfortunately, only a very small part of its 22,937 chapters has survived; these were published in 1963. A number of small encyclopaedias were issued in the 16th century, but the next important event was the publication of the small but profusely illustrated *San ts'ai t'u-hui* (1607–09), compiled by Wang Ch'i and his son Wang Ssu-i. In 1704–11 the Chinese literary encyclopaedia *P'ei-wen yün-fu* was compiled by order of the emperor K'ang-hsi; this was supplemented by the *Yün fu shi I* (1720). Other works ordered by the emperor include the *Pien-tzu lei-pien* (1726) and the *Tzu shih ching hua* (1727). In 1726 the huge *Ku-chin t'u-shu chi-ch'eng* ("Collection of Pictures and Writings") was published by order of the emperor. Edited by the scholar Ch'en Meng-lei, it filled over 750,000 pages and attempted to embody the whole of the Chinese cultural heritage.

At the turn of the century, a number of encyclopaedias were issued. Wang Chi's *Shih wu yüan hui,* which covered well over 2,000 topics, was compiled in 1796. Lu Feng-tso's *Hsiao chih lu* (1804) is particularly valuable for its attention to technical terms, which previous works had ignored. Ch'en Wei's *Ching Chuan II* (1804) concentrated on history and the great Chinese classics, whereas Wang Ch'eng-lieh's *Ch'i ming chi shu* (1806) is stronger in biographical material. Tai Chao-ch'un compiled the *Ssu shu wu ching lei tien chi ch'eng* (1887), a historical work for the use of civil-service candidates. Wei Sung's *I shih chi shih* (1888) had actually been compiled 65 years previously, but it paid far more attention to practical matters. The *Chiu T'ung T'ung* (1902) of Liu K'o-i was in large measure a reassembly of material in the older encyclopaedias in a more efficient classification. A more important work of the period is the largely historical and biographical *Erh shih ssu shih chiu t'ung cheng tien lei yao ho pien* (1902). The *Ch'ing ch'ao hsü wen hsien t'ung k'ao* (1905), compiled by Liu Chin Tsao, was revised and enlarged in 400 volumes in 1921. It includes contemporary material on fiscal, administrative, and industrial affairs and gives some attention to technical matters. Lu Erh-k'uei's *Tz'u-yüan* (1915), with a supplement issued in 1931, was the first really modern Chinese encyclopaedia and set the style for nearly all later works of this nature.

In 1980, officials of the Greater Encyclopedia of China Publishing House and Encyclopædia Britannica, Inc., announced an agreement under which the *Micropædia* of the 15th edition of *Encyclopædia Britannica* would be translated into Chinese for distribution in China. The eight-volume set for this project, named *The Concise Encyclopædia Britannica,* was published serially in 1984–85. It was also announced that an 80-volume set to be known as the *Greater Encyclopedia of China* would be released a volume at a time between 1980 and 1989.

Japan. In the Edo, or Tokugawa, era (1603–1867) there appeared a kind of encyclopaedia that consisted of extracts of major works in Japanese and Chinese. *Kojiruien* (51 volumes, 1879–1914) and *Nihon-hyakka-daijiten,* or the "Great Japanese Encyclopaedia" (10 volumes, 1908–19) were somewhat more akin to modern encyclopaedias but were mostly compilations of scientific works. More complete general encyclopaedias appeared in the Showa period (1926–); *Dai-hyakka* (28 volumes, 1931–35), *Koku-min-hyakka* (15 volumes, 1934–37), *Sekai-daihyakka* (24 volumes, 1955–68), and *Japonica* (19 volumes, 1967–72) are examples of well-compiled works. The *Buritanika Kokusai Dai Hyakka Jiten,* or "*Britannica International Encyclopædia*" (28 volumes), which began publication in 1972 and was completed in 1975, was the joint creation of Encyclopædia Britannica, Inc., and the Tokyo Broadcasting System acting together as TBS/Britannica Company, Tokyo. Unlike most Japanese-language encyclopaedias, which consist largely of simple short entries, its main body consists of 20 volumes of lengthy systematic entries. Other sections of the four-part set include a six-volume reference guide, consisting of many thousands of short factual entries; a reader's guide; and an index. There are also supplemental yearbooks.

The Arab world. The early encyclopaedias written in Arabic can be roughly divided into two classes: those designed for people who wished to be well informed and to make full use of their cultural heritage, and those for the rapidly growing number of official administrators. The latter type of encyclopaedia originated when the Arabs established their rule through so many parts of the Mediterranean region. The first true encyclopaedia was the work of Ibn Qutayba (828–889), a teacher and philologist, who dealt with his topics by quoting traditional aphorisms, historical examples, and old Arabic poems. The arrangement and contents of his *Kitāb ʿUyūn al-Akhbār* ("The Best Traditions") set the pattern for many later encyclopaedias. The 10 books were arranged in the following order: power, war, nobility, character, learning and eloquence, asceticism, friendship, prayers, food, women. Ibn ʿAbd Rabbih of Córdoba improved on Ibn Qutayba's work in his *ʿIqd* ("The Jewelled Necklace") by including more contemporary items of note.

Work of Ibn Qutayba

What has often mistakenly been referred to as the first encyclopaedia, the *Mafātīḥ al-ʿUlūm* ("Key to the Sciences"), was compiled in 975–997 by the Persian scholar and statesman al-Khwārizmī, who was well aware

of the content of the more important Greek writings. He divided his work into two sections: indigenous knowledge (jurisprudence, scholastic philosophy, grammar, secretarial duties, prosody and poetic art, history) and foreign knowledge (philosophy, logic, medicine, arithmetic, geometry, astronomy, music, mechanics, alchemy). The Ikhwān aṣ-Ṣafāʾ ("Sincere Brethren"), a religious or political party founded at Basra in the 10th century, published the *Rasāʾil Ikhwān aṣ-Ṣafāʾ*, a remarkable work that consisted of 52 pamphlets written by five authors, comprising all the knowledge available in their milieu. The work included (1) mathematics, geography, music, logic, and ethics; (2) the natural sciences and philosophy; (3) metaphysics; and (4) religion, astrology, and magic. A complete edition was published in 1887–89.

The Egyptian historian and civil servant an-Nuwairi (1272–1332) compiled one of the best known encyclopaedias of the Mamlūk period, the *Nihāyat al-ʿarab fī funūn al-adab* ("The Aim of the Intelligent in the Art of Letters"), a work of almost 9,000 pages. It comprised: (1) geography, astronomy, meteorology, chronology, geology; (2) man (anatomy, folklore, conduct, politics); (3) zoology; (4) botany; (5) history. A complete edition was issued in 1923. The *Masālik al-abṣār* ("Sight-Seeing Journeys") of al-ʿUmarī (1301–48) was chiefly strong on history, geography, and poetry. A third Egyptian, al-Kalka-shandī (died 1418), compiled a more important and well-organized encyclopaedia, *Ṣubḥ al-aʿshā*, that covered geography, political history, natural history, zoology, mineralogy, cosmography, and time measurement. Ibshīhī (flourished 1440) compiled a very individual encyclopaedia, the *Mustatraf* ("Spiritual Discoveries"), that covered the Islāmic religion, conduct, law, spiritual qualities, work, natural history, music, food, and medicine. At the turn of the Arab fortunes, Ibshīhī had recapitulated all that was best in their culture.

15th- and 16th-century Arabic encyclopaedias

The Persian lawyer ad-Dauwānī (1427–1501) published a kind of encyclopaedia, entitled *Unmūdag al-ʿūlūm*, that consisted of documented questions and answers and technical inventions on a very wide range of subjects. As-Sīrazī (died 1542) soon issued a refutation to it, the *Maqālatar radd ʿalā unmūdag al-ʿulum al-Galā līja*. The *Magmaʾ multaqat az-zuhūr birauḍa min al-manẓūm wal mantur* (1524) of al-Ḥanafī comprised an encyclopaedic survey and description of the various branches of knowledge, with an appendix containing an alphabetical list of the names of God. In Lebanon, Buṭrus al-Bustānī and his sons compiled the *Dāʾirat al-maʿārif* (1876–1900). A second edition (1923–25) was prepared by Muḥammad Farid Wajdi, and a third edition was begun by Fuʾād Afrām al-Bustānī in 1956. In 1955 Albert Rihani issued the one-volume *al-Mawsūʿat al-ʾarabīyah*.

Other areas. Other important encyclopaedias from the East include:

Burmese: the Burma Translation Society's *Myanma swezon kyan/Encyclopaedia Burmanica* (1954–).

Hebrew: *ha-Entziqlopedia ha-ʿIvrit* (*Encyclopaedia Hebraica* 1949–); *Entziqlopedia tevel u-melo'a* (1962–); *Entziqlopedia kelalit Massadah* (1958–60).

Indonesian: *Ensiklopedia Indonesia* (1954).

Sinhalese: *Sinhalese Encyclopedia* (1963–).

(R.L.C./W.E.P.)

DICTIONARIES

The distinction between a dictionary and an encyclopaedia is easy to state but difficult to carry out in a practical way: a dictionary explains words, whereas an encyclopaedia explains things. Because words achieve their usefulness by reference to things, however, it is difficult to construct a dictionary without considerable attention to the objects and abstractions designated. Nonetheless, while a modern encyclopaedia may still be called a dictionary, no good dictionary has ever been called an encyclopaedia.

Historical background

FROM CLASSICAL TIMES TO 1604

In the long perspective of human evolutionary development, dictionaries have been known through only a slight fraction of language history. People at first simply talked without having any authoritative backing from reference books. A short Akkadian word list, from central Mesopotamia, has survived from the 7th century BC. The Western tradition of dictionary making began among the Greeks, although not until the language had changed so much that explanations and commentaries were needed. After a 1st-century-AD lexicon by Pamphilus of Alexandria, many lexicons were compiled in Greek, the most important being those of the Atticists in the 2nd century, that of Hesychius of Alexandria in the 5th century, and that of Photius and the *Suda* in the Middle Ages. (The Atticists were compilers of lists of words and phrases thought to be in accord with the usage of the Athenians.)

Because Latin was a much-used language of great prestige well into modern times, its monumental dictionaries were important and later influenced English lexicography. In the 1st century BC, Marcus Terentius Varro wrote a treatise *De lingua Latina;* the extant books of its section of etymology are valuable for their citations from Latin poets. At least five medieval scholastics—Papias the Lombard, Alexander Neckam, Johannes de Garlandia (John Garland), Hugo of Pisa, and Giovanni Balbi of Genoa—turned their attention to dictionaries. The mammoth work of Ambrogio Calepino, published at Reggio (now Reggio nell'Emilia), in 1502, incorporating several other languages besides Latin, was so popular that "calepin" came to be an ordinary word for a dictionary. A Lancashire will of 1568 contained the provision: "I wyll that Henry Marrecrofte shall have my calapyne and my parafrasies." This is an early instance of the tendency that, several centuries later, caused people to say, "Look in Johnson" or "Look in Webster."

Interlingual dictionaries

Because language problems within a single language do not loom so large to ordinary people as those that arise in the learning of a different language, the interlingual dictionaries developed early and had great importance. The corporation records of Boston, Lincolnshire, have the following entry for the year 1578:

> That a dictionarye shall be bought for the scollers of the Free Scoole, and the same boke to be tyed in a cheyne, and set upon a deske in the scoole, whereunto any scoller may have accesse, as occasion shall serve.

The origin of the bilingual lists can be traced to a practice of the early Middle Ages, that of writing interlinear glosses—explanations of difficult words—in manuscripts. It is but a step for these glosses to be collected together at the back of a manuscript and then for the various lists—glossaries—to be assembled in another manuscript. Some of these have survived from the 7th and 8th centuries—and in some cases they preserve the earliest recorded forms in English.

The first bilingual glossary to find its way into print was a French–English vocabulary for the use of travellers, printed in England by William Caxton without a title page, in 1480. The words and expressions appeared in parallel columns on 26 leaves. Next came a Latin–English vocabulary by a noted grammarian, John Stanbridge, published by Richard Pynson in 1496 and reprinted frequently. But far more substantial in character was an English–Latin vocabulary called the *Promptorius puerorum* ("Storehouse [of words] for Children") brought out by Pynson in 1499. It is better known under its later title of *Promptorium parvulorum sive clericorum* ("Storehouse for Children or Clerics") commonly attributed to Geoffrey the Grammarian (Galfridus Grammaticus), a Dominican friar of Norfolk, who is thought to have composed it about 1440.

The next important dictionary to be published was an English–French one by John (or Jehan) Palsgrave in 1530, *Lesclaircissement de la langue francoise* ("Elucidation of the French Tongue"). Palsgrave was a tutor of French

in London, and a letter has survived showing that he arranged with his printer that no copy should be sold without his permission,

> lest his proffit by teaching the Frenche tonge myght be mynished by the sale of the same to suche persons as, besids hym, wern disposed to studye the sayd tongue.

A Welsh–English dictionary by William Salesbury in 1547 brought another language into requisition: *A Dictionary in Englyshe and Welshe moche necessary to all suche Welshemen as wil spedly learne the Englyshe tôgue.* The encouragement of Henry VIII was responsible for an important Latin–English dictionary that appeared in 1538 from the hand of Sir Thomas Elyot. Thomas Cooper enlarged it in subsequent editions and in 1565 brought out a new work based upon it—*Thesaurus Linguae Romanae et Britannicae* ("Thesaurus of the Roman Tongue and the British"). A hundred years later John Aubrey, in *Brief Lives,* recorded Cooper's misfortune while compiling it:

Thomas Cooper and his *Thesaurus*

> His wife . . . was irreconcileably angrie with him for sitting-up late at night so, compileing his Dictionary. . . . When he had halfe-donne it, she had the opportunity to gett into his studie, tooke all his paines out in her lap, and threw it into the fire, and burnt it. Well, for all that, that good man had so great a zeale for the advancement of learning, that he began it again, and went through with it to that perfection that he hath left it to us, a most usefull worke.

More important still was Richard Huloet's work of 1552, *Abecedarium Anglo-Latinum,* for it contained a greater number of English words than had before appeared in any similar dictionary. In 1556 appeared the first edition by John Withals of *A shorte Dictionarie for Yonge Beginners,* which gained greater circulation (to judge by the frequency of editions) than any other book of its kind. Many other lexicographers contributed to the development of dictionaries. Certain dictionaries were more ambitious and included a number of languages, such as John Baret's work of 1573, *An Alvearie: or triple Dictionarie, in Englishe, Latin, and French.* In his preface Baret acknowledged that the work was brought together by his students in the course of their exercises, and the title *Alvearie* was to commemorate their "beehive" of industry.

The first rhyming dictionary

The first rhyming dictionary, by Peter Levens, was produced in 1570—*Manipulus Vocabulorum. A Dictionarie of English and Latine wordes, set forthe in suche order, as none heretofore hath ben.*

The interlingual dictionaries had a far greater stock of English words than were to be found in the earliest all-English dictionaries, and the compilers of the English dictionaries, strangely enough, never took full advantage of these sources. It may be surmised, however, that people in general sometimes consulted the interlingual dictionaries for the English vocabulary. The anonymous author of *The Arte of English Poesie,* thought to be George Puttenham, wrote, in 1589, concerning the adoption of southern speech as the standard:

> herein we are already ruled by th' English Dictionaries and other bookes written by learned men, and therefore is needeth none other direction in that behalfe.

The mainstream of English lexicography is the word list explained in English. The first known English–English glossary grew out of the desire of the supporters of the Reformation that even the most humble Englishman should be able to understand the Scriptures. William Tyndale, when he printed the Pentateuch on the Continent in 1530, included "A table expoundinge certeyne wordes." The following entries are typical:

> Albe, a longe garment of white lynen.
> Boothe, an housse made of bowes.
> Brestlappe or brestflappe, is soche a flappe as thou seist in the brest or a cope.
> Consecrate, to apoynte a thinge to holy uses.
> Dedicate, purifie or sanctifie.
> Firmament: the skyes.
> Slyme was . . . a fattenesse that osed out of the erth lyke unto tarre/And thou mayst call it cement/if thou wilt.
> Tabernacle, an house made tentwise, or as a pauelion.
> Vapor/a dewymiste/as the smoke of a sethynge pott.

Spelling reformers long had a deep interest in producing English dictionaries. In 1569 one such reformer, John Hart, lamented that the "disorders and confusions" of spelling were so great that "there can be made no perfite Dictionarie nor Grammer." But a few years later the phonetician William Bullokar promised to produce such a work and stated, "A dictionary and grammar may stay our speech in a perfect use for euer."

The schoolmasters also had a strong interest in the development of dictionaries. In 1582 Richard Mulcaster, of the Merchant Taylors' school and later of St. Paul's, expressed the wish that some learned and laborious man "wold gather all the words which we vse in our English tung," and in his book commonly referred to as *The Elementarie* he listed about 8,000 words, without definitions, in a section called "The Generall Table." Another schoolmaster, Edmund Coote, of Bury St. Edmund's, in 1596 brought out *The Englishe Scholemaister, teachinge all his schollars of what age soever the most easie short & perfect order of distinct readinge & true writinge our Englishe tonge,* with a table that consisted of about 1,400 words, sorted out by different typefaces on the basis of etymology. This is important, because what is known as the "first" English dictionary, eight years later, was merely an adaptation and enlargement of Coote's table.

FROM 1604 TO 1828

First purely English dictionary

In 1604 at London appeared the first purely English dictionary to be issued as a separate work, entitled *A Table Alphabeticall, conteyning and teaching the true writing and understanding of hard usuall English wordes, borrowed from the Hebrew, Greeke, Latine,* or *French &c.,* by Robert Cawdrey, who had been a schoolmaster at Oakham, Rutland, about 1580, and in 1604 was living at Coventry. He had the collaboration of his son Thomas, a schoolmaster in London. This work contained about 3,000 words but was so dependent upon three sources that it can rightly be called a plagiarism. The basic outline was taken over from Coote's work of 1596, with 87 percent of his word list adopted. Further material was taken from the Latin–English dictionary by Thomas Thomas, *Dictionarium linguae Latinae et Anglicanae* (1588). But the third source is most remarkable. In 1599 a Dutchman known only as A.M. translated from Latin into English a famous medical work by Oswald Gabelkhouer, *The Boock of Physicke,* published at Dort, in the Netherlands. As he had been away from England for many years and had forgotten much of his English, A.M. sometimes merely put English endings on Latin words. When friends told him that Englishmen would not understand them, he compiled a list of them, explained by a simpler synonym, and put it at the end of the book. Samples are: "Puluerisated, reade beaten; Frigifye, reade coole; Madefye, reade dipp; Calefye, reade heat; Circumligate, reade binde; Ebulliated, read boyled." Thus, the fumblings of a Dutchman who knew little English (in fact, his errata) were poured into Cawdrey's word list. But other editions of Cawdrey were called for—a second in 1609, a third in 1613, and a fourth in 1617.

The next dictionary, by John Bullokar, *An English Expositor,* is first heard of on May 25, 1610, when it was entered in the Stationers' Register (which established the printer's right to it), but it was not printed until six years later. Bullokar introduced many archaisms, marked with a star ("onely used of some ancient writers, and now growne out of use"), such as "aye," "eld," "enewed," "fremd," "gab," and "glee." The work had 14 editions, the last as late as 1731.

Still in the tradition of hard words was the next work, in 1623, by Henry Cockeram, the first to have the word dictionary in its title: *The English Dictionarie: or, an Interpreter of hard English Words.* It added many words that have never appeared anywhere else—adpugne, adstupiate, bulbitate, catillate, fraxate, nixious, prodigity, vitulate, and so on. Much fuller than its predecessors was Thomas Blount's work of 1656, *Glossographia: or, a dictionary Interpreting all such hard words . . . as are now used in our refined English tongue.* He made an important forward step in lexicographical method by collecting words from his own reading that had given him trouble; and he often cited the source. Much of Blount's material was appropriated two years later by Edward Phillips, a nephew of the

poet Milton, for a work called *The New World of English Words,* and Blount castigated him bitterly.

Kersey's *New English Dictionary*

Thus far, the English lexicographers had all been men who made dictionaries in their leisure time or as an avocation, but in 1702 appeared a work by the first professional lexicographer, John Kersey the Younger. This work, *A New English Dictionary,* incorporated much from the tradition of spelling books and discarded most of the fantastic words that had beguiled earlier lexicographers. As a result, it served the reasonable needs of ordinary users of the language. Kersey later produced some bigger works, but all of these were superseded in the 1720s, when Nathan Bailey, a schoolmaster in Stepney, issued several innovative works. In 1721 he produced *An universal etymological English Dictionary,* which for the rest of the century was more popular even than Dr. Johnson's. A supplement in 1727 was the first dictionary to mark accents for pronunciation. Bailey's imposing *Dictionarium Britannicum* of 1730 was used by Samuel Johnson as a repository during the compilation of the monumental dictionary of 1755.

Many literary men felt the inadequacy of English dictionaries, particularly in view of the continental examples. The Accademia della Crusca, of Florence, founded in 1582, brought out its *Vocabolario* at Venice in 1612, filled with copious quotations from Italian literature. The Académie Française produced its dictionary in 1694, but two other French dictionaries were actually more scholarly—that of César-Pierre Richelet in 1680 and that of Antoine Furetière in 1690. In Spain the Royal Spanish Academy (Real Academia Española), founded in 1713, produced its *Diccionario de la lengua Castellana,* 1726–39, in six thick volumes. The foundation work of German lexicography, by Johann Leonhard Frisch, *Teutsch-Lateinisches Wörterbuch,* in 1741, freely incorporated quotations in German. The Russian Academy of Arts (St. Petersburg) published the first edition of its dictionary somewhat later, from 1789 to 1794. Both the French and the Russian academies arranged the first editions of their dictionaries in etymological order but changed to alphabetical order in the second editions.

Samuel Johnson's *Plan*

In England, in 1707, the antiquary Humphrey Wanley set down in a list of "good books wanted," which he hoped the Society of Antiquaries would undertake: "A dictionary for fixing the English language, as the French and Italian." A number of noted authors made plans to fulfill this aim (Joseph Addison, Ambrose Philips, Alexander Pope, and others), but it remained for a promising poet and critic, Samuel Johnson, to bring such a project to fulfillment. Five leading booksellers of London banded together to support his undertaking, and a contract was signed on June 18, 1746. Next year Johnson's *Plan* was printed, a prospectus of 34 pages, consisting of a discussion of language that can still be read as a masterpiece in its judicious consideration of linguistic problems.

With the aid of six amanuenses to copy quotations, Johnson read widely in the literature up to his time and gathered the central word-stock of the English language. He included about 43,500 words (a few more than the number in Bailey), but they were much better selected and represented the keen judgment of a man of letters. He was sympathetic to the desire of that age to "fix" the language, but he realized as he went ahead that "language is the work of man, of a being from whom permanence and stability cannot be derived." At most, he felt that he could curb "the lust for innovation."

The chief glory of Johnson's dictionary was its 118,000 illustrative quotations. No doubt some of these were included for their beauty, but mostly they served as the basis for his sense discriminations. No previous lexicographer had the temerity to divide the verb "take," transitive, into 113 senses and the intransitive into 21 more. The definitions often have a quaint ring to modern readers because the science of the age was either not well developed or was not available to him. But mostly the definitions show a sturdy common sense, except when Johnson used long words sportively. His etymologies reflect the state of philology in his age. Usually they were an improvement on those of his predecessors, because he had as a guide the *Etymologicum Anglicanum* of Franciscus Junius, as edited by Edward Lye, which became available in 1743 and which provided guidance for the important Germanic element of the language.

OA'TMEAL. *n. ſ.* [*oat* and *meal.*] Flower made by grinding oats.

Oatmeal and butter, outwardly applied, dry the ſcab on the head. *Arbuthnot on Aliment.*

Our neighbours tell me oft, in joking talk,
Of aſhes, leather, *oatmeal,* bran, and chalk. *Gay.*

OA'TMEAL. *n. ſ.* An herb. *Ainſworth.*

OATS. *n. ſ.* [æten, Saxon.] A grain, which in England is generally given to horſes, but in Scotland ſupports the people.

It is of the graſs leaved tribe; the flowers have no petals, and are diſpoſed in a looſe panicle: the grain is eatable. The meal makes tolerable good bread. *Miller.*

The *oats* have eaten the horſes. *Shakeſpeare.*

It is bare mechaniſm, no otherwiſe produced than the turning of a wild *oatbeard,* by the inſinuation of the particles of moiſture. *Locke.*

For your lean cattle, fodder them with barley ſtraw firſt, and the *oat* ſtraw laſt. *Mortimer's Huſbandry.*

His horſe's allowance of *oats* and beans, was greater than the journey required. *Swift.*

OA'TTHISTLE. *n. ſ.* [*oat* and *thiſtle.*] An herb. *Ainſ.*

The definition of "Oats" (top) by Samuel Johnson in his *Dictionary* of 1755 exemplifies his prejudice against the Scots and shows his divergence from his source, Nathan Bailey (bottom), who interspersed idiomatic examples throughout his entries (1736).

By courtesy of the Newberry Library, Chicago

OARS, [oran. *Sax.* aora *Su.*] a boat for carrying paſſengers, with two men to row it; alſo inſtruments wherewith boats are row'd.

To have an OAR in every Man's Boat.

That is, to meddle with every man's concerns,

OATS [of æten or etan. *Sax.* to eat] a grain, food for horſes.

To ſow one's wild OATS.

That is to play one's youthful pranks.

OAT *Thiſtle,* an herb.

OA'TEN, of or pertaining to oats

OATH [aŏ, *Sax.* Eeð, *Dan* and *Su.* Eedt, *Da.* Eyd, G.] a ſwearing, or confirming a thing by ſwearing.

OATH [in a legal ſenſe] a ſolemn action, whereby God is called to witneſs the truth of an affirmation, given before one or more perſons impowered to receive the ſame.

OAT-MEAL [of æten and mealepe, *Sax.*] meal or flour made of oats.

Four editions of the *Dictionary* were issued during Dr. Johnson's lifetime; in particular the fourth, in 1773, received much personal care in revision. The *Dictionary* retained its supremacy for many decades and received lavish, although not universal, praise; some would-be rivals were bitter in criticism. A widely heralded work of the 1780s and 1790s was the projected dictionary of Herbert Croft, in a manuscript of 200 quarto volumes, that was to be called the *Oxford English Dictionary.* Croft was, however, unable to get it into print.

Pronouncing dictionaries

The practice of marking word stress was taken over from the spelling books by Bailey in his *Dictionary* of 1727, but a full-fledged pronouncing dictionary was not produced until 1757, by James Buchanan; his was followed by those of William Kenrick (1773), William Perry (1775), Thomas Sheridan (1780), and John Walker (1791), whose decisions were regarded as authoritative, especially in the United States.

The attention to dictionaries was thoroughly established in U.S. schools in the 18th century. Benjamin Franklin, in 1751, in his pamphlet "Idea of the English School," said, "Each boy should have an English dictionary to help him over difficulties." The master of an English grammar school in New York in 1771, Hugh Hughes, announced: "Every one of this Class will have Johnson's Dictionary

in Octavo." These were imported from England, because the earliest dictionary printed in the U.S. was in 1788, when Isaiah Thomas of Worcester, Massachusetts, issued an edition of Perry's *Royal Standard English Dictionary.* The first dictionary compiled in America was *A School Dictionary* by Samuel Johnson, Jr. (not a pen name), printed in New Haven, Connecticut, in 1798. Another, by Caleb Alexander, was called *The Columbian Dictionary of the English Language* (1800) and on the title page claimed that "many new words, peculiar to the United States," were inserted. It received abuse from critics who were not yet ready for the inclusion of American words.

In spite of such attitudes, Noah Webster, already well known for his spelling books and political essays, embarked on a program of compiling three dictionaries of different sizes that included Americanisms. In his announcement on June 4, 1800, he entitled the largest one *A Dictionary of the American Language.* He brought out his small dictionary for schools, the *Compendious,* in 1806 but then engaged in a long course of research into the relation of languages, in order to strengthen his etymologies. At last, in 1828, at the age of 70, he published his master work, in two thick volumes, with the title *An American Dictionary of the English Language.* His change of title reflects his growing conservatism and his recognition of the fundamental unity of the English language. His selection of the word list and his well-phrased definitions made his work superior to previous works, although he did not give illustrative quotations but merely cited the names of authors. The dictionary's worth was recognized, although Webster himself was always at the centre of a whirlpool of controversy.

SINCE 1828

New trends in dictionary making

It was Noah Webster's misfortune to be superseded in his philology in the very decade that his masterpiece came out. He had spent many years in compiling a laborious "Synopsis" of 20 languages, but he lacked an awareness of the systematic relationships in the Indo-European family of languages. Germanic scholars such as Jacob Grimm, Franz Bopp, and Rasmus Rask had developed a rigorous science of "comparative philology," and a new era of dictionary making was called for. Even as early as 1812 Franz Passow had published an essay in which he set forth the canons of a new lexicography, stressing the importance of the use of quotations arranged chronologically in order to exhibit the history of each word. The brothers Jacob and Wilhelm Grimm developed these theories in their preparations for the *Deutsches Wörterbuch* in 1838. The first part of it was printed in 1852, but the end was not reached until more than a century later, in 1960. French scholarship was worthily represented by Maximilien-Paul-Émile Littré, who began working on his *Dictionnaire de la langue française* in 1844, but, with interruptions of the Revolution of 1848 and his philosophical studies, he did not complete it until 1873.

Among scholars in England the historical outlook took an important step forward in 1808 in the work of John Jamieson on the language of Scotland. Because he did not need to consider the "classical purity" of the language, he included quotations of humble origin; in his *Etymological Dictionary of the Scottish Language,* his use of "mean" sources marked a turning point in the history of lexicography. Even as late as 1835 the critic Richard Garnett said that "the only good English dictionary we possess is Dr. Jamieson's Scottish one." Another collector, James Jermyn, showed by his publications between 1815 and 1848 that he had the largest body of quotations assembled before that of *The Oxford English Dictionary.* Charles Richardson was also an industrious collector, presenting his dictionary, from 1818 on, distributed alphabetically throughout the *Encyclopaedia Metropolitana* (vol. 14 to 25) and then reissued as a separate work in 1835–37. Richardson was a disciple of the benighted John Horne Tooke, whose 18th-century theories long held back the development of philology in England. Richardson excoriated Noah Webster for ignoring "the learned elders of lexicography" such as John Minsheu (whose *Guide into the Tongues* appeared in 1617), Gerhard Johannes Vossius (who published his *Etymologicum linguae Latinae* in 1662), and Franciscus Junius (*Etymologicum Anglicanum,* written before 1677). Richardson did collect a rich body of illustrative quotations, sometimes letting them show the meaning without a definition, but his work was largely a monument of misguided industry that met with the neglect it deserved.

The beginnings of the *OED*

Scholars more and more felt the need for a full historical dictionary that would display the English language in accordance with the most rigorous scientific principles of lexicography. The Philological Society, founded in 1842, established an Unregistered Words Committee," but, upon hearing two papers by Richard Chenevix Trench in 1857—"On Some Deficiencies in Our English Dictionaries"—the society changed its plan to the making of *A New English Dictionary on Historical Principles.* Forward steps were taken under two editors, Herbert Coleridge and Frederick James Furnivall, until, in 1879, James Augustus Henry Murray, a Scot known for his brilliance in philology, was engaged as editor. A small army of voluntary readers were inspirited to contribute quotation slips, which reached the number of 5,000,000 in 1898, and no doubt 1,000,000 were added after that. Only 1,827,306 of them were used in print. The copy started going to the printer in 1882; Part I was finished in 1884. Later, three other editors were added, each editing independently with his own staff—Henry Bradley, from the north of England, in 1888, William Alexander Craigie, another Scot, in 1901, and Charles Talbut Onions, the only "Southerner," in 1914. So painstaking was the work that it was not finished until 1928, in over 15,500 pages with three long columns each. An extraordinary high standard was maintained throughout. The work was reprinted, with a supplement, in 12 volumes in 1933 with the title *The Oxford English Dictionary,* and as the *OED* it has been known ever since.

In the United States, lexicographical activity has been unceasing since 1828. In the middle years of the 19th century, a "war of the dictionaries" was carried on between the supporters of Noah Webster and those of his rival, Joseph Emerson Worcester. To a large extent, this was a competition between publishers who wished to preempt the market in the lower schools, but literary people took sides on the basis of other issues. In particular, the contentious Noah Webster had gained a reputation as a reformer of spelling and a champion of American innovations, while the quiet Worcester followed traditions.

In 1846 Worcester brought out an important new work, *A Universal and Critical Dictionary of the English Language,* which included many neologisms of the time, and in the next year Webster's son-in-law, Chauncey Allen Goodrich, edited an improved *American Dictionary* of the deceased Webster. In this edition the Webster interests were taken over by an aggressive publishing firm, the G. & C. Merriam Company. Their agents were very active in the "war of the dictionaries" and sometimes secured an order, by decree of a state legislature, for their book to be placed in every schoolhouse of the state. Worcester's climactic edition of 1860, *A Dictionary of the English Language,* gave him the edge in the "war," and James Russell Lowell declared: "From this long conflict Dr. Worcester has unquestionably come off victorious." The Merriams, however, brought out their answer in 1864, popularly called "the unabridged," with etymologies supplied by a famous German scholar, Karl August Friedrich Mahn. Thereafter, the Worcester series received no major re-editing, and its faltering publishers allowed it to pass into history.

The Century Dictionary

One of the best English dictionaries ever compiled was issued in 24 parts from 1889 to 1891 as *The Century Dictionary,* edited by William Dwight Whitney. It contained much encyclopaedic material but bears comparison even with the *OED.* Isaac Kauffman Funk, in 1893, brought out *A Standard Dictionary of the English Language,* its chief innovation being the giving of definitions in the order of their importance, not the historical order. Thus, at the turn of the new century, the U.S. had four reputable dictionaries—Webster's, Worcester's (already becoming moribund), the *Century,* and Funk's *Standard.* England was also well served by many (the original dates given here)—John Ogilvie (1850), P. Austin Nuttall (1855), Robert Gor-

don Latham (1866, re-editing Todd's Johnson of 1818), Robert Hunter (1879), and Charles Annandale (1882).

Kinds of dictionaries

GENERAL-PURPOSE DICTIONARIES

Although one may speak of a "general-purpose" dictionary, it must be realized that every dictionary is compiled with a particular set of users in mind. In turn, the public has come to expect certain conventional features (see below *Features and problems*), and a publisher departs from the conventions at his peril. One of the chief demands is that a dictionary should be "authoritative," but the word authoritative is ambiguous. It can refer to the quality of scholarship, the employment of the soundest information available, or it can describe a prescriptive demand for compliance to particular standards. Many people ask for arbitrary decisions in usage choices, but most linguists feel that, when a dictionary goes beyond its function of recording accurate information on the state of the language, it becomes a bad dictionary.

Most people encounter dictionaries in the abridged sizes, commonly called "desk" or "college-size" dictionaries. Such handy abridgments go back to the 18th century; Dr. Johnson issued an octavo size in 1756. Their form had become stultified until, in the 1930s, Edward Lee Thorndike, drawing upon the principles of the psychology of learning, produced a series for schools (*Beginning, Junior,* and *Senior*). His dictionaries were not "museums" but tools that encouraged schoolchildren to learn about language. He drew upon his word counts and his "semantic counts" to determine inclusions. The new mode was carried on to the college level by Clarence L. Barnhart in *The American College Dictionary (ACD),* in 1947, and in the later college-size works that were revised to meet that competition—the Merriam-Webster *Seventh New Collegiate* (1963), the *Standard College Dictionary* (1963), and *Webster's New World Dictionary* (1953, and second edition 1970). An especially valuable addition was *The Random House Dictionary* (1966), edited by Jess Stein in a middle size called "the unabridged" and by Laurence Urdang in a smaller size (1968). The Merriam-Webster *Collegiate* series was subsequently extended to eighth (1973) and ninth (1983) editions.

The *Third New International*

The Merriam-Webster *New International* of 1909 had a serene, uncluttered air that suited a simpler age. The second edition, completely reedited, appeared in 1934, and it, in turn, was superseded in 1961 by the *Third New International,* edited by Philip Babcock Gove. Because its competitors of similar size have not been kept up to date, it stands alone among American dictionaries in giving a full report on the lexicon of present-day English. Unfortunately, the advance publicity, before publication, emphasized the quotations from ephemeral writers such as Polly Adler, Ethel Merman, and Mickey Spillane and the statement about "ain't" as "used orally in most parts of the U.S. by many cultivated speakers." Such reports aroused a storm of denunciation in newspapers and magazines by writers who, others asserted, revealed a shocking ignorance of the nature of language. The comments were collected in a "casebook" entitled *Dictionaries and That Dictionary,* edited by James H. Sledd and Wilma R. Ebbitt (1962).

In 1969 came *The American Heritage Dictionary,* edited by William Morris, who was known for his valuable small dictionary *Words* (1947). The *American Heritage* was designed to take advantage of the reaction against the Merriam-Webster *Third.* A "usage panel" of 104 members, chosen mostly from the conservative "literary establishment," provided material for a set of "usage notes." Their pronouncements, found by scholars to be inconsistent, were supposed to provide "the essential dimension of guidance," as the editor put it, "in these permissive times." The etymological material was superior to that in comparable dictionaries.

In England, Henry Cecil Wyld produced his *Universal Dictionary of the English Language* (1932), admirable in every way except for its social class elitism. The smaller sized dictionaries of the Oxford University Press deserve their wide circulation.

SCHOLARLY DICTIONARIES

Thesaurus Linguae Latinae

Beyond the dictionaries intended for practical use are the scholarly dictionaries, with the scientific goal of completeness and rigour in their chosen area. Probably the most scholarly dictionary in the world is the *Thesaurus Linguae Latinae,* being edited in Germany. Its main collections were made from 1883 to 1900, but by 1969 its publication had reached only the letter *O.* A number of countries have "national dictionaries" under way—projects that often take many decades. Two have already been mentioned—the Grimm dictionary for German (a new edition begun in 1965) and the Littré for French (reedited 1956–58); but, in addition, there are the *Woordenboek der Nederlandsche taal* for Dutch, begun in 1882 and now very close to completion; the *Ordbok öfver svenska språket,* for Swedish, begun in 1882, reaching *S* in 1965; the *Slovar sovremennogo russkogo literaturnogo yazika* ("Dictionary of Modern Literary Russian," begun in 1950); the *Nynorsk ordbok* projected for Norwegian; and *Det Norske litteraere ordboksverk* projected for Danish. Of outstanding scholarship are the *Dictionary of Sanskrit on Historical Principles* being prepared at Pune (Poona), India, and *The Historical Dictionary of the Hebrew Language,* now getting under way in Jerusalem. The most ambitious project of all is located at the Centre National in Nancy, France, directed by Paul Imbs, preparing for a *Trésor de la langue française.* In the decade following 1960, over 250,000,000 word examples were collected, the latest techniques of computerization being used. It remains to be seen how much of this can be printed. A laboratory at Besançon, under the direction of B. Quemada, for contemporary French, has extensive collections.

The *Oxford English Dictionary* remains as the supreme completed achievement in all lexicography. Its size makes its revision impractical and the decision was therefore made for supplementation rather than revision. In 1919 plans were pushed forward for a set of "period dictionaries." After the completion of the *OED* in 1928, the remaining quotations, both used and unused, were divided up for use in each project. The prime mover of this plan, Sir William Craigie, undertook *A Dictionary of the Older Scottish Tongue* himself, covering the period from the 14th to the 17th century in Scottish speech. Enough material was amassed under his direction so that editing could begin in 1925, and before his death in 1957 he arranged that it should be carried on at the University of Edinburgh. By 1971 it had reached the word "natural." The work on the older period spurred the establishment of a project on modern Scots, which got under way in 1925, called *The Scottish National Dictionary,* giving historical quotations after the year 1700. By 1971 the project had passed the word "stane."

Period dictionaries for English

For the mainstream of English, a period dictionary for Old English (before 1100) was planned for many decades by a dictionary committee of the Modern Language Association of America (Old English section), but only in the late 1960s did it get under way at the Pontifical Institute of Mediaeval Studies at the University of Toronto. Plans are for the dictionary to be based on a combining of computerized concordances of bodies of Old English literature. A *Middle English Dictionary* has fared much better, convering the period 1100 to 1475. Started in 1925, it had reached the middle *L*'s by 1971, with an overwhelming fullness of detail. For the period 1475 to 1700, an *Early Modern English Dictionary* has not fared as well. It got under way in 1928 at the University of Michigan, and over 3,000,000 quotation slips were amassed, but the work could not be continued in the decade of the Great Depression, and only in the middle 1960s was it revived again. The *OED* supplement of 1933 is again being supplemented—this time in three large volumes, the first of which was published in 1972.

Craigie, in 1925, proposed a dictionary of American English. Support was found for the project, and he transferred from Oxford University to the University of Chicago in order to become its editor. The aim of the work, he wrote, was that of "exhibiting clearly those features by which the English of the American colonies and the United States is distinguished from that of England and the rest of the

English-speaking world." Thus, not only specific Americanisms were dealt with but words that were important in the natural history and cultural history of the New World. After a 10-year period of collecting, publication began in 1936 under the title *A Dictionary of American English on Historical Principles,* and the 20 parts (four volumes) were completed in 1944. This was followed in 1951 by a work that limited itself to Americanisms only—*A Dictionary of Americanisms,* edited by Mitford M. Mathews.

The English language, as it has spread widely over the world, has come to consist of a group of coordinate branches, each expressing the needs of its speakers in communication; further scholarly dictionaries are needed to record the particular characteristics of each branch. Both Canada and Jamaica were treated in 1967—*A Dictionary of Canadianisms on Historical Principles,* Walter Spencer Avis, editor in chief, and *Dictionary of Jamaican English,* edited by Frederic G. Cassidy and R.B. LePage. A historical dictionary of South African English is under way at Rhodes University, Grahamstown, South Africa, edited by William Branford, and some day full dictionaries must be compiled for Australian English, New Zealand English, and so on. Such dictionaries are valuable in displaying the intimate interrelations of the language to the culture of which it is a part.

SPECIALIZED DICTIONARIES

Earliest English etymological dictionary

Specialized dictionaries are overwhelming in their variety and their diversity. Each area of lexical study, such as etymology, pronunciation, and usage, can have a dictionary of its own. The earliest important dictionary of etymology for English was Stephen Skinner's *Etymologicon Linguae Anglicanae* of 1671, in Latin, with a strong bias for finding a classical origin for every English word. In the 18th century, a number of dictionaries were published that traced most English words to Celtic sources, because the authors did not realize that the words had been borrowed into Celtic rather than the other way around. With the rise of a soundly based philology by the middle of the 19th century, a scientific etymological dictionary could be compiled, and this was provided in 1879 by Walter William Skeat. It has been kept in print in re-editions ever since but was superseded in 1966 by *The Oxford Dictionary of English Etymology,* by Charles Talbut Onions, who had worked many decades on it until his death. Valuable in its particular restricted area is J.F. Bense's *Dictionary of the Low-Dutch Element in the English Vocabulary* (1926–39).

Two works are especially useful in showing the relation between languages descended from the ancestral Indo-European language—Carl Darling Buck's *Dictionary of Selected Synonyms in the Principal Indo-European Languages* (1949) and Julius Pokorny's *Indogermanisches etymologisches Wörterbuch* (1959). The Indo-European roots are well displayed in the summary by Calvert Watkins, published as an appendix to *The American Heritage Dictionary* mentioned earlier. Interrelations are also dealt with by Eric Partridge in his *Origins* (1958).

The pronouncing dictionary, a type handed down from the 18th century, is best known in the present day by two examples, one in England and one in America. That of Daniel Jones, *An English Pronouncing Dictionary,* represents what is "most usually heard in everyday speech in the families of Southern English persons whose men-folk have been educated at the great public boarding-schools." Although he called this the Received Pronunciation (RP), he had no intention of imposing it on the English-speaking world. It originally appeared in 1917 and was repeatedly revised during the author's long life. Also strictly descriptive was a similar U.S. work by John S. Kenyon and Thomas A. Knott, *A Pronouncing Dictionary of American English,* published in 1944 and never revised but still valuable for its record of the practices of its time.

The "conceptual dictionary," in which words are arranged in groups by their meaning, had its first important exponent in Bishop John Wilkins, whose *Essay towards a Real Character and a Philosophical Language* was published in 1668. A plan of this sort was carried out by Peter Marc Roget with his *Thesaurus,* published in 1852 and many times reprinted and re-edited. Although philosophically oriented, Roget's work has served the practical purpose of another genre, the dictionary of synonyms.

Dictionaries of usage

The dictionaries of usage record information about the choices that a speaker must make among rival forms. In origin, they developed from the lists of errors that were popular in the 18th century. Many of them are still strongly puristic in tendency, supporting the urge for "standardizing" the language. The work with the most loyal following is Henry Watson Fowler's *Dictionary of Modern English Usage* (1926), ably re-edited in 1965 by Sir Ernest Gowers. It represents the good taste of a sensitive, urbane litterateur. It has many devotees in the U.S. and also a number of competitors. Among the latter, the most competently done is *A Dictionary of Contemporary American Usage* (1957), by Bergen Evans and Cornelia Evans. Usually the dictionaries of usage have reflected the idiosyncrasies of the compilers; but, from the 1920s to the 1960s, a body of studies by scholars emphasized an objective survey of what is in actual use, and these were drawn upon by Margaret M. Bryant for her book *Current American Usage* (1962). A small corner of the field of usage is dealt with by Eric Partridge in *A Dictionary of Clichés* (1940).

The regional variation of language has yielded dialect dictionaries in all the major languages of the world. In England, after John Ray's issuance of his first glossary of dialect words in 1674, much collecting was done, especially in the 19th century under the auspices of the English Dialect Society. This collecting culminated in the splendid *English Dialect Dictionary* of Joseph Wright in six volumes (1898–1905). American regional speech was collected from 1774 onward; John Pickering first put a glossary of Americanisms into a separate book in 1816. The American Dialect Society, founded in 1889, made extensive collections, with plans for a dictionary, but this came to fruition only in 1965, when Frederic G. Cassidy embarked on *A Dictionary of American Regional English* (known as *DARE*).

The many "functional varieties" of English also have their dictionaries. Slang and cant in particular have been collected in England since 1565, but the first important work was published in 1785, by Capt. Francis Grose, *A Classical Dictionary of the Vulgar Tongue,* reflecting well the low life of the 18th century. In 1859 John Camden Hotten published the 19th-century material, but a full historical, scholarly survey was presented by John Stephen Farmer and W.E. Henley in their *Slang and Its Analogues,* in seven volumes, 1890–1904, with a revised first volume in 1909 (all reprinted in 1971). For the present century, the dictionaries of Eric Partridge are valuable. Slang in the United States is so rich and varied that collectors have as yet only scratched the surface, but the work by Harold Wentworth and Stuart B. Flexner, *Dictionary of American Slang* (1960), can be consulted. The argot of the underworld has been treated in many studies by David W. Maurer.

Bilingual dictionaries

Of all specialized dictionaries, the bilingual group are the most serviceable and frequently used. With the rise of the vernacular languages during the Renaissance, translating to and from Latin had great importance. The Welshman in England was provided with a bilingual dictionary as early as 1547, by William Salesbury. Scholars in their analyses of language, as well as practical people for everyday needs, are anxious to have bilingual dictionaries. Even the most exotic and remote languages have been tackled, often by religious missionaries with the motive of translating the Bible. The finding of exact equivalents is more difficult than is commonly realized, because every language slices up the world in its own particular way.

Dictionaries dealing with special areas of vocabulary are so overwhelming in number that they can merely be alluded to here. In English, the earliest was a glossary of law terms published in 1527 by John Rastell. His purpose, he said, was "to expown certeyn obscure & derke termys concernynge the lawes of thys realme." The dictionaries of technical terms in many fields often have the purpose of standardizing the terminology; this normative aim is especially important in newly developing countries where the language has not yet become accommodated to modern

technological needs. In some fields, such as philosophy, religion, or linguistics, the terminology is closely tied to a particular school of thought or the individual system of one writer, and, consequently, a lexicographer is obliged to say, "according to Kant," "in the usage of Christian Science," "as used by Bloomfield," and so on.

Features and problems

ESTABLISHMENT OF THE WORD LIST

Problems in selecting a word list

The goal of the big dictionaries is to make a complete inventory of a language, recording every word that can be found. The obsolete and archaic words must be included from the earlier stages of the language and even the words attested to only once (nonce words). In a language with a large literature, many "uncollected words" are likely to remain, lurking in out-of-the-way sources. The *OED* caught many personal coinages, but not "head-over-heelishness" (1882), "odditude" (1860), "pigstyosity" (1869), "whitechokerism" (1866), and other graceless jocularities. Also, the so-called latent words are a problem, when a lexicographer knows that a derivative word probably has been used, but he has no evidence for it. The *OED* had three quotations for "kindheartedness" but none for "kindheartedly," which any speaker of English would feel free to use. Some "ghost words" have arisen from the misreading of manuscripts and from misprints, and the lexicographer attempts to cast these out.

Various large blocks of words have a questionable status. Both geographic names and biographical entries are selectively included in some dictionaries but are really encyclopaedic. More than 2,000,000 insects have been identified and named by entomologists, while names of chemical compounds and drugs may be almost as numerous. Trade names and proprietary names may number in the hundreds of thousands. Vogue suffixes like "-ism," "-ology," "-scope," or "-wise" are used by some with the freedom of a grammatical construction. These millions are beyond what any dictionary can be expected to include.

For the smaller-sized dictionaries, the editors attempt to choose the words that are likely to be looked up. They comb the scholarly works carefully and supplement them from files that they may have collected. They may decide to put derivative words at the end of entries as "run-ons" or to have all words strictly as separate alphabetical entries. The size is ultimately decided by the commercial consideration of how much can be put into a work that can be sold for a reasonable price and held readily in the hand. (Bulk also influences the size of the word list for unabridged dictionaries.)

The establishment of a word list involves many difficult technical problems. Linguists tend to use the terms morpheme, free form, bound form, lexeme, and so on, inasmuch as "word" is a popular term not suited to technical use. A safe compromise is to use "lexical unit." This term allows the inclusion of set phrases (established groups) and idioms. Words having different etymological sources must be considered as different words. Thus "calf" in the sense of the young of a bovine animal came from Common Germanic, whereas "calf" for the fleshy back of the lower part of the leg came from Old Norse, perhaps from a Celtic source. A more difficult problem is found when a word entered the language at different points—such as "cookie," from the Dutch *koekje* "little cake," recorded in Scottish in 1701 in the form *cuckie,* then independently taken from the Dutch of the Hudson Valley in the form *cockie* in 1703, and perhaps independently taken into South African English from Afrikaans in the mid-19th century.

SPELLING

British and American spelling

Dictionaries have probably played an important role in establishing the conventions of English spelling. Dr. Johnson has received much credit for this, though he differed very little from his predecessors. He used the spelling "smoak" in the early part of his dictionary, but when he came to the entry itself, he changed it to "smoke," and this has prevailed. Noah Webster introduced some simplifications that have become accepted in American English. American dictionaries usually label the distinctive British spellings, such as "centre" and its class, "honour" and its class, "connexion," "gaol," "kerb," "tyre," "waggon," and a few others.

The desire for uniformity is so great that popular variants are not welcomed; the very common "alright" is not yet approved, nor is the widespread "miniscule" for "minuscule." The *OED* is exceptional in listing the early variant spellings, showing that a common word like "good" has been spelled in 13 different ways, with seven more from Scottish usage. When the spelling reform movement was at its height, from the 1880s to about 1910, the dictionaries included the new forms, but in recent years these have been expunged. The graphic dress of the language is now so sacrosanct that dictionaries are used as authoritarian "style manuals" in matters of spelling, hyphenation, and syllabification.

PRONUNCIATION

Dictionaries are more responsive to usage in the matter of pronunciation than they are in spelling. It is claimed that in the 19th century the Merriam-Webster dictionaries foisted a New England pronunciation on the United States, but in recent years many regional variations have been recorded. *Webster's Third New International* (1961) went to surprising lengths in its variants; perhaps its record is in giving 132 different ways of pronouncing "a fortiori."

The former practice of giving pronunciations as if the words were pronounced in isolation in a formal manner represented an artificiality that distorted language in use; recent dictionaries have marked pronunciation as it appears in continuous discourse. Furthermore, there has been a trend toward what has been called "democratization." In the word "government," for instance, it is recognized that many people do not pronounce an *n,* and some people actually say something like "gubb-munt." There is a constant battle between traditional spoken forms and spelling pronunciations.

Systems for indicating sounds

Since the alphabet is notoriously inadequate for recording the sounds of English, dictionaries are forced to adopt additional symbols. A system of using numerals over vowels was handed down from the 18th century, but that gave way to the diacritic markings of the Merriam-Webster series. The rise of the International Phonetic Alphabet (IPA) has offered another possibility, but the general public as yet finds it abstruse. Even more detailed symbols are needed in linguistic atlases and phonetic research. With considerable courage, Clarence L. Barnhart introduced the symbol schwa (ə) into *The American College Dictionary* (1947) for the neutral midcentral vowel, as at the beginning and end of "America," and the symbol has now become widely accepted. Although some systems are clumsier than others, the key does not matter much if it is applied consistently.

ETYMOLOGY

The supplying of etymologies involves such difficult decisions for a lexicographer as whether words should be carried back into prehistory by means of reconstructed forms or the degree to which speculation should be permitted. A U.S. Romance scholar, Yakov Malkiel, has presented the notion that words follow "trajectories"—by finding certain points in the history of a word, one can link up the developments in form and meaning. The austere treatment of some words consists in saying "derivation unknown," and yet this sometimes causes interesting possibilities to be ignored.

Native stock and loanwords

A fundamental distinction is made in word history between the "native stock" and the "loanwords." There have been so many borrowings into English that the language has been called "hypertrophied." The traditional view is to regard the borrowings as a source of "richness." A historical dictionary does its best to ascertain the date at which a word was adopted from another language, but the word may have to go through a period of probation. Murray, the editor of the *OED,* listed four stages of word "citizenship": the casual, the alien, the denizen, and the natural. The casuals may not be part of the language, as they appear only in travel writings and accounts of foreign

countries, but a lexicographer must collect citations for them in order to record the early history of a word that may later become naturalized. Some words may remain "denizens" for centuries, Murray pointed out, such as "phenomenon" treated as Greek, "genus" as Latin, and "aide-de-camp" as French. When a word is borrowed, its etymology may be traced through its descent in its original language.

Some early philosophies have assumed that there is a mystic relation between the present use of a word and its origin and that etymology is a search for the "true meaning." The recognition of continuous linguistic change establishes, however, that etymology is no more than early history, sometimes as reconstructed on the basis of relationships and known sound changes. Ingenuity in etymologizing is dangerous, and even plausibility can be misleading, but ascertained fact has overriding importance. It is curious that recent slang is often more uncertain in its origin than words of long history.

GRAMMATICAL INFORMATION

Dictionaries are obliged to contain the two basic kinds of words of a language—the "function words" (those that perform the grammatical functions in a language, such as the articles, pronouns, prepositions, and conjunctions) and the "referential words" (those that symbolize entities outside the language system). Each kind must be treated in a suitable way. Dictionaries have been much criticized for not including a sufficiency of grammatical information. It is usual to mark the part of speech, but not the categories of mass noun and count noun. (A mass noun, such as "milk" or "oxygen," cannot ordinarily be used in the plural, while a count noun is any noun that can be pluralized.) Such information is given in some dictionaries designed for teaching, and the technique could well be adopted more generally. The irregular inflections must be given, showing that one says "goose," "geese," but not "moose," "meese." Or in the verbs one says "walk," "walked," but "ride," "rode." It is usual to treat the different parts of speech as separate lexical entries, as in "to walk" and "to take a walk," requiring a parallel list of senses, but Edward Lee Thorndike, in his school dictionaries, experimented with grouping the parts of speech together when they had a similar sense.

Grammar and vocabulary

The relation of grammar to the vocabulary is the subject of considerable controversy among linguists. If one considers the analysis of language as one unified enterprise, then the grammar is central and the lexical units are inserted at some point in the analysis. Another view is that the division is into coordinate branches, such as phonology, syntax, and lexicon. Certainly lexicographers try to take advantage of all findings made by grammarians.

SENSE DIVISION AND DEFINITION

A language like English has so many complex developments in the senses—*i.e.*, the particular meanings—of its words that the task of the lexicographer is difficult. It is generally accepted that "meaning" is a suffusing characteristic of all language by definition, and the attempt to slice meaning into "senses" must be done arbitrarily by the person analyzing the language. This is where collected contexts form the basis of the lexicographer's judgment. He sorts the quotations into piles on the basis of similarities and differences and he may have to discard "transitional" examples. Figurative developments, such as the "mouth" of a river or the "foot" of a hill, make complications in the relationships.

For the order in which the senses of words are given, the order of historical development has been chiefly used. For an old word like "earth," the information may be insufficient. The editors of the *OED* had to give up, because, they said, "Men's notions of the shape and position of the earth have so greatly changed since Old Teutonic times"; they were obliged to compromise with a logical order. Sometimes, but not always, a word seems to have a "core," or central, meaning from which other meanings develop. If the historical order is followed, the obsolete and archaic meanings may have to appear first; and, therefore, some popular dictionaries give the most important meaning first and work down to the rare and occasional meanings at the end. The so-called "semantic count," giving senses in order of frequency, has also been used.

There seems to be no one method that is best for defining all words. The lexicographer must use artistry in selecting the ways that will convey a sense accurately and succinctly. He attempts to find what is "criterial" in a particular meaning, but he can also give further detail until he runs into the area of the encyclopaedic.

In logical theory it would be ideal to have a "metalanguage" in which definitions could be stated, but nothing of the sort is available for popular use. A "defining vocabulary" can be established, and in school dictionaries the definitions use simple words. In the last analysis all definitions have to fall back on undefined terms (to be accepted like axioms) that symbolize first-order experience of life. In this connection the logician Willard Quine has argued that lexicography is basically concerned with synonymy.

USAGE LABELS

Variations in language use

Part of the information that a dictionary should give concerns the restrictions and constraints on the use of words, commonly called usage labelling. There is great variation in language use in many dimensions—temporal, geographical, and cultural. The people who make a two-part division into "correct" and "incorrect" show that they do not understand how language works. The valuation does not lie in the word itself but in the appropriateness of the context. Therefore, it is preferable to be sparing in the use of labels and to allow the tone to become apparent from the illustrative examples. An important distinction was put forward in 1948 by an American philologist, John S. Kenyon, when he discriminated between "cultural levels," which refer to the degree of education and cultivation of a person, and "functional varieties," which refer to the styles of speech suitable to particular situations. Thus a cultivated person rightly uses informal or colloquial language when at ease with friends.

A lexicographer is faced with the difficult task of selecting a suitable set of labels. In the temporal categories, labels such as obsolete, obsolescent, archaic, and old-fashioned are dangerous, because some speakers have long memories and might use old words very naturally. The national labels are problematical, because words move easily from one branch of the language to another. The word "blizzard," for instance, is no doubt an Americanism in origin, but, since the 1880s, it has been so well known over the English-speaking world that a national label would be misleading. The label "dialect" or "regional," either for England or America, offers many problems, for alleged "boundaries" are permeable. The label "colloquial" was much misunderstood, and now "informal" is often used in its place. There may be a "poetic vocabulary" that needs labelling, and few people will agree on any definition of "slang."

It is revealing that in early printings of the Merriam-Webster *Third New International* under the word "cock-eyed," marked "slang," one of the quotations is by a careful stylist named Jacques Barzun; in order to use effective English, as he does, this cultivated writer is willing to draw upon slang. Some would argue that in marking the use as "slang," the Merriam-Webster staff was not sufficiently "permissive."

Some dictionaries wisely include special paragraphs on the constraints of usage, sometimes as a "synonymy" and sometimes as a "usage note."

ILLUSTRATIVE QUOTATIONS

Citation of sources

Dictionaries of the past have copied shamelessly from one to another, but the collecting of a file of illustrative quotations makes possible a fresh, original treatment. Scholarly works like the *OED* and its supplementations follow the canon of always using the earliest quotation and the latest for an obsolete word; in between, the quotations are selected for revealing facets of usage or for "forcing" a meaning. The criterion of use by only the best writers does not hold for a truly historical dictionary, because a "low" source may be especially revealing. The giving of exact source citations is not a matter of pedantry but es-

tablishes the scientific basis by which others can check the evidence. A different set of quotations, accurately attested, might have led to a different treatment. Thus the phrase "illustrative quotation" is something of a misnomer, for the quotations are more than "illustrative"; they form the basic evidence from which conclusions are drawn. It is the work of the editor to decide when the collections are sufficient—"ripe," as it were—to move from the collecting stage to the editing stage.

A small-sized dictionary may advantageously use made-up sentences, because an aptly framed "forcing" context can tell more than a definition. In fact, the habitual collocations of a word (the surrounding words with which it usually appears) may be revealing of the nature of a word. "Dictionaries of collocations" may be a step forward in future lexicography.

TECHNOLOGICAL AIDS

The development of machine aids, such as the computer, has been heralded by some as ushering in a new era in lexicography. Although the computer can do well in many tasks of great drudgery—mechanical excerpting of texts, alphabetizing, and classifying by designated descriptors—it is limited to what a human being tells it to do. It is difficult for a computer to sort out homographs—separate words that are spelled alike; and, at the editing stage, the delicate decisions must be humanly made.

Uses for computers

The computer can be used to good advantage in the compilation of concordances of individual authors or of limited texts, and then one type of dictionary could be made by a summation of concordances. Such a procedure, with a large body of literature like that of the Renaissance, would overwhelm an editor. More feasible, perhaps, is the establishment of a computerized archive that would never be published, but would serve as a storehouse from which, by advanced retrieval methods, the desired information could be called forth at will. The *Trésor de la langue française* of Nancy, already mentioned, is a step in this direction.

ATTITUDES OF SOCIETY

Without a doubt, dictionaries have been a conservative force for many hundreds of years, not only in countries that have had an official academy that has the national language as part of its province but also in the English-speaking countries, in which academies have been spurned. Well-entrenched popular attitudes account for this. A Neoplatonic outlook assumes that there exists an ideal form of language from which faltering human beings have departed and that dictionaries might bring people closer to the perfect language. Also, there is a widespread "yearning for certainty," a seeking for guidance amid the wilderness of possible forms. Thus, people welcome self-proclaimed "supreme authorities."

Americans have had additional reasons for their homage to the dictionary. In colonial times Americans felt themselves to be far from the centre of civilization and were willing to accept a book standard in order to learn what they thought prevailed in England. This linguistic colonialism lasted a long time and set the pattern of accepting the dictionary as a "lawgiver." In 1869, a cultural leader declared: "Upon the proper spelling, pronunciation, etymology, and definition of words, a dictionary might be made to which high and almost absolute authority might justly be awarded." In this vein teachers have taken pains to inculcate "the dictionary habit" in their pupils. Rather than observe the language around them, as Englishmen commonly do, Americans give up their autonomy and fly to a dictionary to settle questions on language. This call for dogmatic prescription has been a source of uneasiness to lexicographers, most of whom now argue that all they can do legitimately is to describe how the language has been used.

Observance of taboos

Social attitudes have affected the dictionaries also in the enforcement of certain taboos. Certain words commonly called obscene have been omitted, and, thus, irrational taboos have been strengthened. If the sex words were given in their alphabetical place, with suitable labels, the false attitudes in society would more readily be cleansed. A perennial problem in lexicography is the treatment of the terms of ethnic insult, such as "dago," "kike," and "wop." There is constant social pressure for leaving them out, and some dictionaries have succumbed to it, but it may be that an enlightened attitude shows that the open discussion of prejudices is the best way of getting rid of them.

The greatest value of a dictionary is in giving access to the full resources of a language and as a source of information that will enhance free enjoyment of the mother tongue.

Major dictionaries

For the English language the important dictionaries have already been cited in the appropriate sections; but the supreme achievement represented by the *OED* should be emphasized again. The major dictionaries in some other languages may be mentioned here.

Dictionaries of the Romance languages

For the French language, the Académie's dictionary is now in its eighth edition (1931–35) and manifests conservative views about the vocabulary, but three other works are actually more serviceable—the *Petit Larousse: dictionnaire encyclopédique pour tous* (1959); a new edition of the famous Littré, *Dictionnaire de la langue française* (1974), four volumes; and a splendid new work, Paul Robert, *Dictionnaire alphabétique et analogique de la langue française* (1960–64), six volumes. For French etymology alone, the standard work is Walther von Wartburg, *Französisches etymologisches Wörterbuch,* nearing completion in 1970 in volume 18, with a few gaps to be filled.

Among other Romance tongues, Italian has had many dictionaries. The Accademia della Crusca of Florence furnished its *Vocabolario* in a first edition in 1612, but the edition begun in 1863 bogged down at the letter *O* in 1923, and a successor work, begun in 1941, has not gone far. There is also the dictionary by G. Devoto and G.C. Oli, *Dizionario della lingua italiana* (1971). Following the model of the *OED* is the still uncompleted *Grande dizionario della lingua italiana* (1961), edited by Salvatore Battaglia. Very serviceable to English speakers is the *Italian Dictionary* of Alfred Hoare (1915; second edition, 1925) and that of Barbara Reynolds, begun in 1962, and still under way. For Spanish, the Real Academia Española in Madrid has done well since its first edition in 1726–39. At present the 18th edition, from 1956, is available. Contributions from New World Spanish need further scholarly treatment.

For the German language, the great dictionary begun by the brothers Grimm, completed in 1960, is to be re-edited in a project that will take many years; but, meanwhile, a standard work is that of Hermann Paul, *Deutsches Wörterbuch,* which first appeared in 1897 but is now available in a sixth edition (1968). The national dictionaries in the Scandinavian countries were mentioned above, but a work done with special scholarly skill is noteworthy: Einar Haugen, editor in chief, *Norwegian English Dictionary* (Madison, Wisconsin [Oslo printed], 1965), dealing with the two official languages of Norway, Bokmål and Nynorsk. Another form of a Germanic language, Afrikaans, which developed from the Dutch transplanted to South Africa in the 17th century, has a big dictionary under way. Publication of *Woordeboek van die Afrikaanse taal* began at Pretoria in 1950 as a collaboration of the best scholars in South Africa. A full dictionary of Yiddish requires profound scholarship, and this was provided by Uriel Weinreich in *Modern English-Yiddish, Yiddish-English Dictionary* (1968).

Special problems of Greek dictionaries

Greek lexicography offers special difficulties because of the long range of illustrious literature that must be covered and the split in recent centuries between Katharevusa, the literary language, and Demotic, the language of everyday life. For the English-speaking world, the standard work for Ancient Greek is by Henry George Liddell and Robert Scott, *A Greek-English Lexicon,* published in a first edition in 1843, but now available in a ninth edition, 1925–40. A full dictionary of Demotic, edited by Demetrius Georgacas at the University of North Dakota, is still in the project stage. For Russian the Soviet Academy of Arts has produced a useful work in four volumes (1957–61), but a more detailed one has been in progress since 1950, reach-

ing *F* in 1964. The Royal Irish Academy is at work on a definitive dictionary of Irish, but only "contributions" and certain parts are so far available. Many linguists have attempted to cover Arabic; probably the most useful work is that of Hans Wehr, as translated and edited by J. Milton Cowan, *A Dictionary of Modern Written Arabic* (1961). For Japanese, the standard source is the *Dai-jiten* ("Great Dictionary"), issued at Tokyo in 26 volumes (1934–36). The best known Chinese dictionary, *Tz'u hai,* was revised in 1969 and published in Taipei, Taiwan.

Titles of other works are to be found in the bibliographies listed below, especially in Constance Winchell's *Guide.*

(A.W.Re.)

BIBLIOGRAPHY

Encyclopaedias: There are two short and very readable introductions to the subject: LIBRARY OF CONGRESS, *The Circle of Knowledge* (1979), a well-illustrated guide issued in connection with a Library of Congress exhibition; and SIGFRID H. STEINBERG, "Encyclopaedias," *Signature,* New Series, no. 12, pp. 3–22 (1951), which is a brilliant conspectus of the whole field of encyclopaedia history. ROBERT L. COLLISON, *Encyclopaedias: Their History Throughout the Ages,* 2nd ed. (1966), lists and describes in one chronological sequence encyclopaedias from both East and West, and pays particular attention to *L'Encyclopédie, Brockhaus,* the *Britannica,* the *Metropolitana,* and *Larousse.* It also includes a reprint (pp. 238–295) of SAMUEL TAYLOR COLERIDGE's philosophical essay on the design of encyclopaedias, the "Treatise on Method." FRITZ SAXL, "Illustrated Mediaeval Encyclopaedias," in his *Lectures,* vol. 1, pp. 228–254, and vol. 2, plates 155–174 (1957, reprinted 1978), is an important and original contribution to the subject, the 20 illustrations being especially interesting. The *Journal of World History* devoted one complete issue (vol. 9, no. 3, 1966) to an international symposium on encyclopaedias, special attention being paid to St. Isidore, Hugh of Saint-Victor, Raoul Ardent, Vincent of Beauvais, Sahagún, *L'Encyclopédie,* the *Metropolitana,* the *Britannica, L'Encyclopédie française,* and Arabic and Chinese encyclopaedias of the classical period. ALBERT J. WALFORD (ed.), *Guide to Reference Material,* 4th ed. (1980–); and EUGENE P. SHEEHY *et al.* (comp.), *Guide to Reference Books,* 9th ed. (1976), and their supplements, both provide scholarly evaluations of the principal current English- and foreign-language encyclopaedias. GERT A. ZISCHKA, *Index Lexicorum: Bibliographie der Lexikalischen Nachschlagewerke* (1959), is important both for its excellent summary of the history of the encyclopaedia and for its extensive bibliography. FRANCES NEEL CHENEY and WILEY J. WILLIAMS, *Fundamental Reference Sources,* 2nd ed. (1980), includes discussions of good encyclopaedias and dictionaries. "The Uses of Encyclopaedias: Past, Present, and Future," in the *American Behavioral Scientist,* 6:3–40 (1962), is a stimulating symposium with contributions by Livio C. Stecchini, Jacques Barzun, Harry S. Ashmore, W.T. Couch, Charles Van Doren, Francis X. Sutton, David L. Sills, Carl F. Stover, and Alfred de Grazia. HERMAN KOGAN, *The Great EB* (1958), is a well-written and fascinating account of the *Britannica* and its history, but it is also valuable for the light it throws on the more practical problems and techniques of the encyclopaedia world in general. S. PADRAIG WALSH, *Anglo-American General Encyclopedias* (1968), is a dictionary of English-language encyclopaedias issued during the period 1703–1967, and has helpful evaluative notes. It has been continued by his *General Encyclopedias in Print, 1973–74: A Comparative Analysis* (1973). AMERICAN LIBRARY ASSOCIATION. REFERENCE AND SUBSCRIPTION BOOKS REVIEW COMMITTEE, *Purchasing a General Encyclopedia* (1969), is a pamphlet suggesting 12 criteria for evaluating the quality and usefulness of any encyclopaedia, and containing the committee's recommendations concerning a number of named encyclopaedias. The ALA's "Six Multivolume Adult Encyclopedias," *Booklist* 79(7):515–532 (December 1, 1982), and "Encyclopedia Yearbooks, Annuals, and Supplements", *Booklist* 77(14):1049–1054 (March 15, 1981), are analytical reviews based on these criteria. BOHDAN S. WYNAR (ed.), *American Reference Books Annual,* a reviewing service for reference books published in the United States, regularly includes overviews of encyclopaedias. KENNETH F. KISTER, *Encyclopedia Buying Guide,* 3rd ed. (1981), is a comprehensive consumer guide to general encyclopaedias in the English language. In each encyclopaedia the entry under the word "Encyclopaedia" or "Encyclopedia" will usually (but not invariably) provide information concerning its own history, and often gives very useful information on the history of encyclopaedias in general. Additional details may often be found in the encyclopaedia's general introduction, which is usually printed in the first volume.

Dictionaries: For the best list of dictionaries, see EUGENE P. SHEEHY *et al.* (comp.) *Guide to Reference Books,* 9th ed. (1976). See also ROBERT L. COLLISON, *Dictionaries of Foreign Languages* (1957); A.J. WALFORD (ed.), *A Guide to Foreign Language Grammars and Dictionaries,* 2nd ed. (1967); WOLFRAM ZAUNMULLER, *Bibliographisches Handbuch der Sprachwörterbücher* (1958); GERT A. ZISCHKA, *Index Lexicorum* (1959); and *Foreign Language-English Dictionaries,* 2 vol. (Library of Congress, Reference Department, 1955). For dictionaries published in Communist countries, see DANUTA RYMSZA-ZALEWSKA (ed.), *Bibliography of Dictionaries* (1965). Dictionaries of Americanisms and of slang are well covered by W.J. BURKE, *The Literature of Slang,* pp. 2–11 (1939). ANNIE M. BREWER, *Dictionaries, Encyclopedias, and Other Word-Related Books,* 3 vol., 3rd ed. (1982), is a classified catalog of more than 28,000 dictionaries and encyclopaedias.

History: For the history of classical dictionaries, see JOHN EDWIN SANDYS, *A History of Classical Scholarship,* 3rd ed., vol. 1, pp. 295–407 (1921). An old standard survey is JAMES A.H. MURRAY, *The Evolution of English Lexicography* (1900). See also MITFORD M. MATHEWS, *A Survey of English Dictionaries* (1933, reprinted 1966); and JAMES ROOT HULBERT, *Dictionaries: British and American,* rev. ed. (1968). For excellent scholarly details in their areas, see DE WITT T. STARNES, *Renaissance Dictionaries: English-Latin and Latin-English* (1954); DE WITT T. STARNES and GERTRUDE E. NOYES, *The English Dictionary from Cawdrey to Johnson, 1604–1755* (1946); and JAMES H. SLEDD and GWIN J. KOLB, *Dr. Johnson's Dictionary: Essays in the Biography of a Book* (1955). For the history of the *Oxford English Dictionary,* see WILLIAM A. CRAIGIE, "Historical Introduction," in the *Supplement* (1933), transferred to vol. 1 in the reissue of 1933; HANS AARSLEFF, "The Early History of the *Oxford English Dictionary," Bulletin of the New York Public Library,* 66:417–439 (1962), and *The Study of Language in England, 1780–1860* (1967), especially ch. 6. For American dictionaries, see JOSEPH HAROLD FRIEND, *The Development of American Lexicography, 1798–1864* (1967). The documents on the controversy over the Merriam-Webster *Third New International Dictionary* are collected by JAMES H. SLEDD and WILMA R. EBBITT in *Dictionaries and That Dictionary* (1962). For discussions of the technical problems arising in lexicography, see FRED W. HOUSEHOLDER and SOL SAPORTA (eds.), *Problems in Lexicography,* 2nd ed. (1967), papers of a conference held in 1960—especially practical is the paper by CLARENCE L. BARNHART, "Problems in Editing Commercial Monolingual Dictionaries," pp. 161–181; LADISLAV ZGUSTA, *Manual of Lexicography* (1971); and ALLEN WALKER READ, "Approaches to Lexicography and Semantics," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 10, pp. 145–205 (1972). An "International Conference on Lexicography in English" was held in New York City, June 5–7, 1972; its proceedings are published in the *Ann. N.Y. Acad. Sci.* (1973). R.R.K. HARTMANN (ed.), *Lexicography: Principles and Practice* (1983), is a collection of papers concerned with the making of dictionaries.

Endocrine Systems

Endocrinology deals with the structure and function of glands that secrete materials internally. It is important to distinguish between an endocrine gland, which discharges substances called hormones directly into the bloodstream or lymph system, and an exocrine gland, which secretes substances through a duct opening in the gland onto an external or internal body surface. Salivary and sweat glands, examples of exocrine glands, secrete saliva and sweat, respectively, which act locally at the site of duct openings. In contrast, hormones that are secreted in minuscule quantities by endocrine glands, are transported by the circulation to exert powerful effects on tissues remote from the site of secretion.

As far back as 3000 BC, the ancient Chinese diagnosed some endocrinologic disorders and were able to provide effective treatments. For example, seaweed, which is rich in iodine, was prescribed for the treatment of goitre (enlargement of the thyroid gland). Perhaps the earliest demonstration in humans of direct endocrinologic intervention was the castration of men who could then be relied upon, more or less, to safeguard the chastity of women living in harems. During the Middle Ages and persisting well into the 19th century, it was a popular practice to castrate pubertal boys to preserve the purity of their treble voices. Castration established the testicle as the source of substances responsible for the development and maintenance of "maleness."

This knowledge led to an abiding interest in restoring or enhancing male sexual powers. John Hunter, an 18th-century Scottish surgeon, anatomist, and physiologist who practiced in London, transplanted successfully the testis (testicle) of a rooster into the abdomen of a hen. Charles-Édouard Brown-Séquard, a 19th-century French neurologist and physiologist, asserted that testes contained an invigorating, rejuvenating substance. His conclusions were based, in part, on observations obtained after he had injected himself with an extract of the testicle of a dog or of a guinea pig to which water had been added. These experiments were advances in that they resulted in the widespread use of organ extracts (organotherapy).

Modern endocrinology, however, is largely a creation of the 20th century. Its scientific origin is firmly rooted in the studies of Claude Bernard (1813–78), a brilliant French physiologist who made the key observation that complex organisms, such as humans, go to great lengths to preserve the constancy of what he called the "milieu intérieur" (internal environment). Later, an American physiologist, Walter Bradford Cannon (1871–1945), used the term homeostasis to describe this inner constancy.

The endocrine system, in association with the nervous system and the immune system, regulates the body's internal activities and external interactions to preserve the static internal environment. This control system permits the prime functions of living organisms—growth, development, and reproduction—to proceed in an orderly, stable fashion; it is exquisitely self-regulating so that any disruption of the normal internal environment by internal or external events is resisted by powerful countermeasures. When this resistance is overcome, sickness ensues.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 421 and 423.

This article is divided into the following sections:

Traditional endocrinology

Because endocrinology involves an actively expanding body of knowledge, its borders remain difficult to define. The traditional core of an endocrine system, however, consists of (1) an endocrine gland, (2) its hormonal secretion, (3) a responding tissue containing a specific receptor to which the hormone will become bound, and (4) the action that results after the hormone becomes bound, termed the postreceptor response.

Embryonic origin of endocrine glands

Each endocrine gland consists of a group of specialized cells that have a common origin in the developing embryo. Many endocrine glands are derived from cells that arise in the embryonic digestive system (*e.g.*, the thyroid and pancreas) or from cells that migrate from the embryonic nervous system (*e.g.*, the parathyroid and adrenal medulla). Still others arise from a region of the embryo known as the urogenital ridge (ovary, testis, and adrenal cortex). The pituitary gland is derived from cells that originate in both the nervous system and the digestive tract. Each endocrine gland has a rich supply of blood, which is directly related to its role in synthesizing and secreting hormones. Many endocrine glands secrete more than one hormone. Some organs function both as exocrine glands and as endocrine glands. The pancreas is the best-known example.

In addition to the more traditional endocrine cells described above, specially modified nerve cells within the nervous system secrete important hormones into the blood. These special nerve cells are called neurosecretory cells, and their secretions are termed neurohormones to distinguish them from the hormones produced by traditional endocrine cells. The areas of the nervous system that produce neurohormones also have a rich vascular supply, and the neurohormones are either released into the blood or stored in adjacent blood-rich areas (called neurohemal organs) until needed.

Most hormones are one of two types: proteins (including peptides and modified amino acids) or steroids. The majority of hormones are the protein type. They are highly soluble in water and can be transported readily through the blood. The protein hormones, when initially synthesized within the cell, are contained within larger, biologically inert molecules called prohormones. The inactive portion of the prohormone is split away so that one or more active fragments that are released from the cell circulate through the blood. A smaller number of hormones are the water-soluble, fatty acid steroid hormones, all of which are synthesized from the precursor molecule cholesterol. These lipid hormones are transported through the blood by first being bound to proteins in the blood.

All body tissues that respond to a specific hormone contain specially shaped molecular configurations called receptors. These receptors are found on the surface of target cells, in the case of protein and peptide hormones, or within the cytoplasm, in the case of steroids and modified amino acid hormones. Each receptor has a strong, highly specific affinity (attraction) for a particular hormone.

This arrangement permits a specific hormone to have an effect only on those tissues for which it is "targeted," namely, those that are equipped with specific binding receptors. Usually, one segment of the hormone molecule exhibits a strong chemical affinity for the receptor while another area is responsible for initiating its specific action. Thus, hormonal actions are not general throughout the body but rather are aimed at specific target tissues.

The hormone-receptor complex

The hormone-receptor complex that is formed then activates a chain of specific chemical responses within the cells of the target tissue to complete the hormonal action. This action may be the result of the activation of enzymes within the target cell, of the interactions of the hormone-receptor complex with the nucleus of the cell, and consequent stimulation of new protein synthesis, or of a combination of both. It may even result in secretion of another hormone.

General features

THE NATURE OF ENDOCRINE REGULATION

Endocrine gland secretion is not a haphazard process; it is subject to precise, intricate control at several levels so that its effect may be integrated with those of the nervous system and the immune system. The simplest level of control resides at the endocrine gland itself. Characteristically, the signal for an endocrine gland to release more or

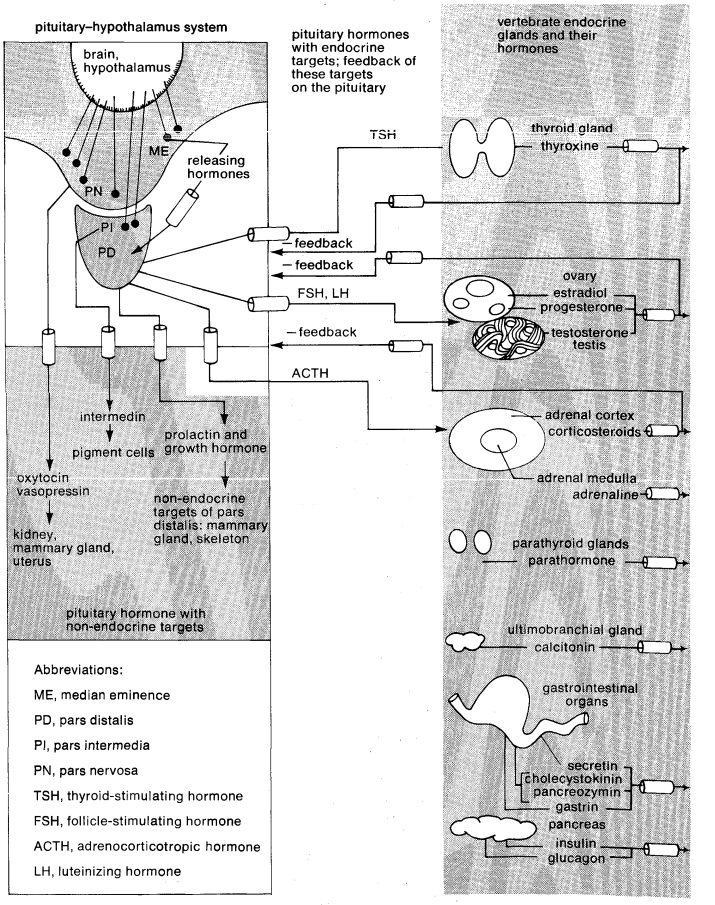


Figure 1: *Relationships of endocrine glands.* At the upper left is the pituitary-hypothalamus axis, and on the upper right are the endocrine target glands controlled through negative feedback by this axis. At the lower left are nonendocrine target organs of this axis. At the lower right are some endocrine glands that are not directly affected by the feedback mechanisms that regulate the endocrine and nonendocrine target organs.

less of its hormone is the level of some substance, either a hormone that influences the action of a gland (called a tropic hormone), a biochemical product such as glucose, or a biologically important element such as potassium or calcium. Because the endocrine gland has a rich supply of blood, it is able to detect changes in the level of this regulating substance.

Simple negative feedback

Some endocrine glands, for example the parathyroid glands located in the neck, are controlled largely by a simple negative feedback mechanism as demonstrated in Figure 2. Parathyroid hormone, known as parathormone (A), acts on its major target organ, bone (B), and other tissues to transport calcium into the blood, raising the serum calcium level (C). Elevated serum calcium levels inhibit the secretion of parathormone by the parathyroid glands (D). Thus, if for any reason serum calcium levels become elevated, parathormone secretion is blocked and calcium is not secreted into the serum from bone; the serum calcium level then falls back into the normal range. If, on the other hand, the serum calcium level should fall (E), the parathyroids are no longer inhibited from releasing parathormone and parathyroid gland activity is stimulated. (F) The increased circulating levels of parathormone stimulate increased dissolution of bone, releasing calcium. Thus, calcium enters into the serum from bone, and the serum calcium concentration rises until it reaches a normal level. In this fashion, in individuals with normal parathyroid glands, serum calcium levels are maintained within a narrow range even in the face of large changes in calcium intake or excessive losses of calcium from the body.

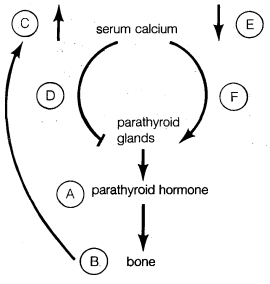


Figure 2: Control of parathyroid hormone secretion (see text).

Control of the hormonal secretions of a number of other endocrine glands is more complex because the glands are, themselves, target organs of a regulatory system called the hypothalamic-pituitary-target organ axis. Glands of this type include the thyroid, the adrenal cortex, and the gonads (testes and ovaries). The major mechanism involves interconnecting negative feedback loops, each similar to that described above, which involve the hypothalamus (a structure located at the base of the brain and above the pituitary), the anterior pituitary, and the target organ. The hypothalamus stimulates the pituitary, through neurohormones, to secrete pituitary hormones, which affect any of a number of target organs. The hypothalamic-pituitary-target organ axis is one of the more elegant devices to be found in nature. A generalized representation is illustrated in Figure 3 and discussed below.

The target gland secretes its hormone (target gland hormone), which (A) combines with the receptors of a secondary target tissue and is then inactivated. This continues until the concentration of target gland hormone in the blood exceeds the amount necessary to bind all of the tissue receptors. The effect of the target gland hormone on the secondary target tissue is quantitative; that is, within limits, the greater the amount of target gland hormone bound to receptors in the secondary target tissue, the greater the secondary target tissue cell response. The target gland hormone also binds to specific receptors in the anterior pituitary (B) to inhibit the secretion of pituitary-stimulating hormone (the hormone that stimulates the target gland to secrete more target gland hormone). As the concentration of the target gland hormone in the blood rises, there is an appropriate decrease in the production of pituitary-stimulating hormone. Thus, there will be less activation (C) of the target gland to produce its hormone. The end result of this feedback mechanism is that the high level of target hormone circulating in the bloodstream falls back to normal.

Conversely, as more target gland hormone is bound (A) to receptors in the secondary target tissue, the levels of target gland hormone circulating in the bloodstream falls. The overall inhibitory effects of target gland hormone on the pituitary gland then is reduced. Low levels of target gland hormone thus stimulate production of more pituitary-stimulating hormone (C), which in turn stimulates the secretion of target gland hormone by the target gland, until (B) the concentration of target gland hormone in the blood increases to a normal level.

A second, similar negative feedback loop is superimposed on the first. The target gland hormone binds to nerve cells in the hypothalamus (D), which inhibit the secretion of specific hypothalamic-releasing hormones (neurohormones) that stimulate the secretion of pituitary hormone (an important element in the previous negative feedback loop). The concentrations of hypothalamic-releasing hormones (E) within a set of veins that connects the hypothalamus and the pituitary gland (the hypophyseal-portal circulation) is reduced.

Nervous influences

The importance of this second loop (D and E) lies in the fact that the nerve cells of the hypothalamus communicate with nervous influences that extend down from the brain (G), including the cerebral cortex (the centre for higher mental functions, movement, perceptions, etc.), thus permitting the endocrine system to respond to physical and emotional stresses. The mechanism involves the interruption of the primary feedback loop (B and C) so that the concentrations of hormones in the blood can be increased or decreased appropriately in response to environmental stresses perceived by the nervous system (see below *The human endocrine system: The hypothalamus*). If this were not available, all blood hormones would be locked in at normal concentrations, even at times when it would be important to the body for these hormones to diverge from normal levels. Similarly, appropriate endocrinologic responses can be achieved from stimuli resulting from signals generated through the immune system to threats (such as bacterial invasion) from within the organism.

Finally, a third short loop (E and F) directly inhibits the release of a specific hypothalamic-releasing hormone by a pituitary hormone. In this fashion, concentrations of pituitary, thyroid, adrenal cortex, and gonadal hormones in the blood are maintained at normal, homeostatic levels, but, when necessary, the hormonal levels may be altered dramatically to meet changing circumstances of the internal or external environment.

This traditional view of the mechanisms that control endocrine secretion has been modified by evidence pointing to important supplemental control mechanisms. When, as is usually the case, more than one cell type is found within a single endocrine gland, the hormones secreted by one cell may exert a direct modulating effect upon the secretions of its immediate neighbour of a different cell type. This form of control is known as paracrine function. Similarly, the secretions of one endocrine cell may affect the activity of a neighbour cell of identical type, an activity known as autocrine function. Thus, endocrine cell activity may be modulated directly from within the endocrine gland itself without the need for hormones to enter the general circulation.

Paracrine and autocrine functions

Excluding from the definition of a hormone the requirement that it act at a site remote from the secreting endocrine cell allows additional classes of bioactive materials to be considered as hormones. Neurotransmitters, a group of chemical compounds of variable composition, are secreted at all synapses (junctions between nerve cells over which nervous impulses must pass). They facilitate or inhibit the transmission of neural impulses and have given rise to the hybrid science of neuroendocrinology (the branch of medicine that studies the interaction of the nervous system and the endocrine system). A second group of novel bioactive substances are called the prostaglandins, a complex group of fatty acids that are formed and secreted

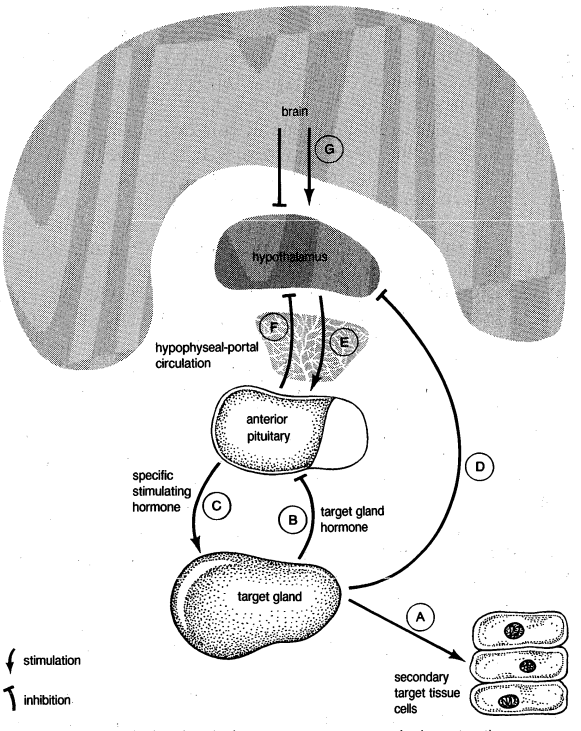


Figure 3: Hypothalamic-pituitary-target organ axis (see text).

by many tissues. They mediate important biological effects in almost every organ system of the body.

Growth factors

Another group of substances with hormonelike actions is called growth factors. These are substances that stimulate the growth of specific target tissue cells. They are distinct from the usual members of the endocrine family of growth hormones in that they were identified only after it was noted that target cells grown outside the organism in tissue culture could be stimulated to grow and reproduce by gland or tissue extracts chemically distinct from any known growth hormone.

Still another area of hormonal classification that has come under intensive investigation is the effect of endocrines on animal behaviour. While simple, direct hormonal effects on human behaviour are difficult to document because of the complexities of human motivation, there are many convincing demonstrations of hormone-mediated behaviour in other life forms. A special case is that of the pheromone, a substance generated by an organism that influences, by its odour, the behaviour of another organism of the same species. An often-quoted example is the musky scent of the females of many species, which provokes sexual excitation in the male. Such devices have obvious adaptive value for species survival.

FUNCTIONS OF THE ENDOCRINE SYSTEM

Maintenance of homeostasis. For an organism to function normally and effectively, it is necessary that the processes of its tissues operate smoothly and conjointly in a stable setting. The endocrine system provides an essential mechanism, called homeostasis, that integrates body activities and at the same time ensures that the composition of the body fluids bathing all of the constituent cells remains constant.

Scientists have postulated that the concentrations of the various salts present in the fluids of the body closely resemble the concentrations of salts in the primordial seas, which nourished the simple organisms from which increasingly complex species have evolved. Since any change in the salt composition in fluids that surround the cells (extracellular fluid), such as the fluid portion of the circulating blood (the intravascular plasma or serum), would necessitate large compensating changes in the salt concentrations within cells, the constancy of these salts (called electrolytes) is closely guarded. Even small changes in the circulating levels of these electrolytes (which include sodium, potassium, chloride, calcium, magnesium, and phosphate) elicit prompt, appropriate responses from the endocrine system, by employing negative feedback regulatory mechanisms similar to those described above, in order to restore normal concentrations.

Homeostasis

Not only is the level of each individual electrolyte maintained through homeostasis, but the total concentration of all of the electrolytes per unit of fluid (called the osmolality) is kept constant as well. If this were not the case, an increase in extracellular osmolality (or an increase in the concentration of electrolytes per unit of fluid) would result in the movement of intracellular fluid out of the cells, across the cell membrane, and into the extracellular compartment. Because the kidneys would excrete much of the fluid from the expanded extracellular volume, dehydration would be the result. Conversely, decreased plasma osmolality (or a decrease in the concentration of electrolytes per unit of fluid) would lead to a buildup of fluid within the cell.

Another homeostatic mechanism involves not an adjustment of the concentration of electrolytes in plasma but, rather, the maintenance of a normal total plasma volume. If the total volume of fluid within the circulation increases (a condition known as overhydration), the pressure against the walls of the blood vessels and the heart increase as well, stimulating sensitive areas in heart and vessel walls to release hormones that modulate the excretion of water and electrolytes by the kidney, thus reducing the total plasma volume to normal.

Growth and differentiation. Despite the many mechanisms designed to maintain a constant internal environment, the organism itself is subject to change; it is born (or hatches), it matures, and it ages. These changes are accompanied by supportive variations in body fluid composition. For example, the normal serum phosphate concentration in a child is about six milligrams per 100 millilitres, whereas about half that value is the normal concentration in an adult. These and other more striking changes are part of a second major function of the endocrine system, namely, the control of normal growth and development. The mammalian fetus develops in the uterus of the mother under the powerful influence not only of hormones from its own endocrine glands but from hormones generated in the mother's placenta as well, a system known as the

Control of growth

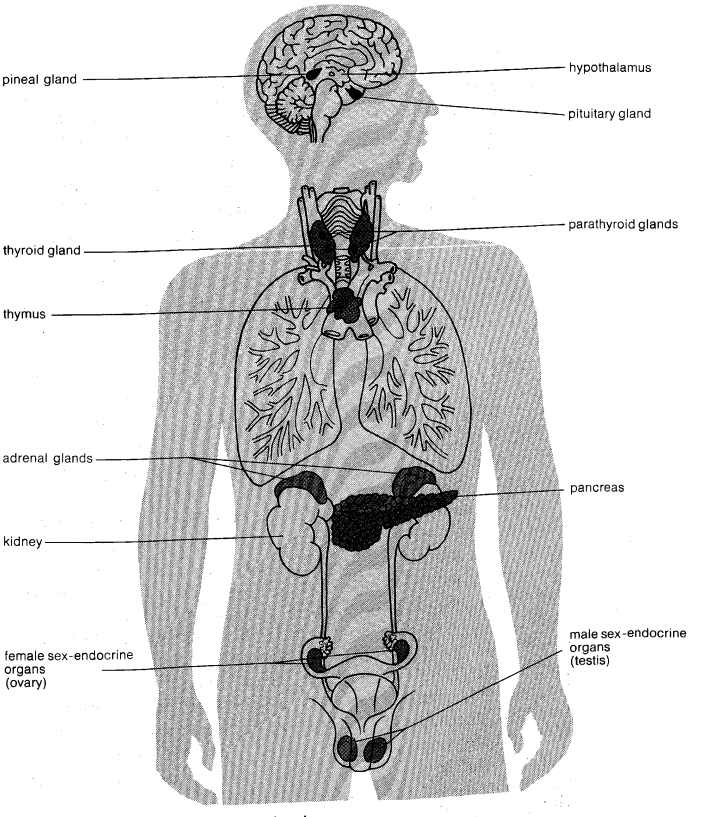


Figure 4: The human endocrine glands.

fetoplacental unit. Maternal endocrine glands assure that a proper mixture of nutrients is transmitted by way of the placenta to the growing fetus. Hormones also are present in the mother's milk and are transferred to the suckling young. There is evidence for the transmission of hormones into the eggs of nonmammals, which may influence the development of the embryo.

Sexual differentiation of the fetus into a male or a female is also controlled by delicately timed hormonal changes. After birth and a period of steady growth in infancy and childhood, the changes associated with puberty and adolescence take place. This dramatic transformation of an adolescent into a physically mature adult is also initiated and controlled by the endocrine system. Finally, as the endocrinology of aging has come under intensive investigation, changes associated with the process of normal aging and senescence have been discovered.

Adaptive responses to stress. Throughout the life of the organism endocrine influences are at play to enhance the ability of the body to respond to internal and external stressful stimuli. These changes allow not only the individual organism but also the species to survive. Early studies by Cannon led him to the thesis that acutely threatened animals respond with multiple physical changes, including endocrine changes, that prepare them to react or retreat, a process known as "the fight or flight reaction."

Adaptive responses for more prolonged stresses also occur. For example, in states of malnutrition typical of the self-induced semistarvation condition called anorexia nervosa, there is reduced secretion of thyroid hormones (hormones that generally stimulate metabolic processes of the body), leading to a lower metabolic rate. This change reduces the rate of the consumption of the body's fuel, and thus reduces the rate of consumption of the remaining energy stores. This change has obvious survival value; death from starvation is deferred.

Parenting behaviour. The endocrine system, particularly the hypothalamus, the anterior pituitary, and the gonads, is intimately involved in reproductive behaviour by providing physical, visual, and olfactory (pheromonal) signals that arouse the sexual interest of the male and the receptivity of the female. Furthermore, there are powerful endocrine influences on parental behaviour in all species, probably including humans.

Pheromones

ANATOMIC CONSIDERATIONS

Figure 4 illustrates those secretory organs that have traditionally made up the human endocrine system. While these obviously glandular structures synthesize and secrete specific hormones (Table 1), studies have revealed that most body tissues may also function as endocrine organs. The growing list includes the lungs, the heart, the skeletal muscles, the uterus, the kidneys, the salivary glands, and the lining of the gastrointestinal tract. Finally, as mentioned above, bundles of nerve cells, called nuclei, have evolved into classical endocrine organs; they secrete neurohormones into the bloodstream. (T.B.S.)

Comparative endocrinology

Comparative endocrinologists investigate the evolution of endocrine systems and the role of these systems in animals' adaptation to their environments and their production of offspring. Studies of nonmammalian animals have provided information that has furthered research in mammalian endocrinology, including that of humans. For example, the actions of a pituitary hormone, prolactin, on the control of body water and salt content were first discovered in fishes and later led to the demonstration of similar mechanisms in mammals. The mediating role of local ovarian secretions (paracrine function) in the maturation of oocytes (eggs) was discovered in starfishes and only later extended to vertebrates. The important role of thyroid hormones during embryonic development was first studied thoroughly in tadpoles during the early 1900s. In addition, the isolation and purification of many mammalian hormones was made possible in large part by using other vertebrates as bioassay systems; that is, primitive animals have served as relatively simple, sensitive indicators of the amount of hormone activity in extracts prepared from mammalian endocrine glands. Finally, some vertebrate and invertebrate animals have provided "model systems" for research that have yielded valuable information on the nature of hormone receptors and the mechanisms of hormone action. For example, one of the most intensively studied systems for understanding hormone actions on target tissues has been the receptors for progesterone and estrogens (hormones secreted by the gonads) from the oviducts of chickens.

Usefulness of comparative studies

Table 1: The Human Endocrine System

gland or tissue	hormone	chemical nature
Testis	testosterone	steroid
Ovary	estrogens (estradiol, estrone, estriol)	steroids
	inhibin (folliculostin)	polypeptide?
	progesterone	steroid
	relaxin	polypeptide
Thyroid gland	thyroxine (T_4)	amino acid
	triiodothyronine (T_3)	amino acid
Adrenal gland		
Medulla	epinephrine	amine
	norepinephrine	amine
	dopamine	steroid
Cortex	cortisol	steroid
	corticosterone	steroid
	aldosterone	steroid
	androgens	steroid
	estrogens	steroid
Pituitary gland		
Anterior lobe	corticotropin (adrenocorticotropin, ACTH)	polypeptide
	growth hormone (GH or somatotropin)	protein
	thyrotropin (thyroid-stimulating hormone, TSH)	glycoprotein
	follicle-stimulating hormone (FSH)	glycoprotein
	luteinizing hormone (LH, interstitial cell stimulating hormone, ICSH)	glycoprotein
	prolactin (PRL, luteotropic hormone, LTH, lactogenic hormone, mammotropin)	protein
Posterior lobe	oxytocin	polypeptide
	vasopressin (antidiuretic hormone, ADH)	polypeptide
Intermediate lobe tissue	α-melanocyte-stimulating* hormone (α-MSH)	polypeptide
	β-melanocyte-stimulating* hormone (β-MSH)	polypeptide
Hypothalamus	corticotropin-releasing hormone (CRH)	polypeptide?
	growth hormone-releasing hormone (GHRH)	polypeptide?
	thyrotropin-releasing hormone (TRH)	polypeptide
	follicle-stimulating hormone (FSH)	polypeptide?
	gonadotropin-releasing hormone (GnRH)	polypeptide
	prolactin-inhibiting factor (PIF)	polypeptide?
	somatostatin	polypeptide
	gastrointestinal neuropeptide	polypeptide
Pancreatic islets	glucagon	polypeptide
	insulin	polypeptide
	somatostatin	polypeptide
	pancreatic polypeptide	polypeptide
Parathyroid gland	parathyroid hormone (parathormone)	polypeptide
	calcitonin	polypeptide
Skin, liver, kidney	calciferols	steroids
Gastrointestinal mucosa		
Stomach	gastrin	polypeptide
Duodenum	cholecystokinin (CCK)	polypeptide
	secretin	polypeptide
	gastric-inhibitory polypeptide (GIP)	polypeptide
	vasoactive intestinal peptide (VIP)	polypeptide
	villikinin	—
	enterocrinin	—
Thymus	thymosin	polypeptide
Pineal	melatonin	amine (not secreted?)
Kidneys	renin	protein
	erythropoietin	protein?
Placenta	human chorionic gonadotropin (HCG)	glycoprotein
	human chorionic somatomammotropin (HCS)	protein
	renin	protein
	estrogens	steroids
	androgens	steroids
	progesterone	steroid
Multiple tissues	somatomedins (insulin-like growth factors)	polypeptides
	prostaglandins	steroids

*Intermediate lobe hormones referred to collectively as melanotropin or intermedin.

An understanding of how the endocrine system is regulated in nonmammals also provides essential information for regulating natural populations or captive animals. Artificial control of salmon reproduction has had important implications for the salmon industry as a whole. Some successful attempts at reducing pest insect species have been based on the knowledge of pheromones. Understanding the endocrinology of a rare species may permit it to be bred successfully in captivity and thus prevent it from becoming extinct. Future research may even lead to the reintroduction of some endangered species into natural habitats.

EVOLUTION OF ENDOCRINE SYSTEMS

Primitive endocrine glands

The most primitive endocrine systems seem to be those of the neurosecretory type, in which the nervous system either secretes neurohormones (hormones that act on, or are secreted by, nervous tissue) directly into the circulation or stores them in neurohemal organs (neurons whose endings directly contact blood vessels, allowing neurohormones to be secreted into the circulation), from which they are released in large amounts as needed. True endocrine glands probably evolved later in the evolutionary history of the animal kingdom as separate, hormone-secreting structures. Some of the cells of these endocrine glands are derived from nerve cells that migrated during the process of evolution from the nervous system to various locations in the body. These independent endocrine glands have been described only in arthropods (where neurohormones are still the dominant type of endocrine messenger) and in vertebrates (where they are best developed).

It has become obvious that many of the hormones previously ascribed only to vertebrates are secreted by invertebrates as well (for example, the pancreatic hormone insulin). Likewise, many invertebrate hormones have been discovered in the tissues of vertebrates, including those of humans. Some of these molecules are even synthesized and employed as chemical regulators, similar to hormones in higher animals, by unicellular animals and plants. Thus, the history of endocrinologic regulators has ancient beginnings, and the major changes that took place during evolution would seem to centre around the uses to which these molecules were put.

VERTEBRATE ENDOCRINE SYSTEMS

Evolutionary relationships

Vertebrates (phylum Vertebrata) are separable into at least seven discrete classes that represent evolutionary groupings of related animals with common features. The class Agnatha, or the jawless fishes, is the most primitive group. Class Chondrichthyes and class Osteichthyes are jawed fishes that had their origins, millions of years ago, with the Agnatha. The Chondrichthyes are the cartilaginous fishes, such as sharks and rays, while the Osteichthyes are the bony fishes. Familiar bony fishes such as goldfish, trout, and bass are members of the most advanced subgroup of bony fishes, the teleosts, which developed lungs and first invaded land. From the teleosts evolved the class Amphibia, which includes frogs and toads. The amphibians gave rise to the class Reptilia, which became more adapted to land and diverged along several evolutionary lines. Among the groups descending from the primitive reptiles were turtles, dinosaurs, crocodilians (alligators, crocodiles), snakes, and lizards. Birds (class Aves) and mammals (class Mammalia) later evolved from separate groups of reptiles. Amphibians, reptiles, birds, and mammals, collectively, are referred to as the tetrapod (four-footed) vertebrates.

The human endocrine system is the product of millions of years of evolution. and it should not be surprising that the endocrine glands and associated hormones of the human endocrine system have their counterparts in the endocrine systems of more primitive vertebrates. By examining these animals it is possible to document the emergence of the hypothalamic-pituitary-target organ axis, as well as many other endocrine glands, during the evolution of fishes that preceded the origin of terrestrial vertebrates.

The hypothalamic-pituitary-target organ axis. The hypothalamic-pituitary-target organ axes of all vertebrates are similar. The hypothalamic neurosecretory system is poorly developed in the most primitive of the living Agnatha vertebrates, the hagfishes, but all of the basic rudiments are present in the closely related lampreys. In most of the more advanced jawed fishes there are several well-developed neurosecretory centres (nuclei) in the hypothalamus that produce neurohormones. These centres become more clearly defined and increase in the number of distinct nuclei as amphibians and reptiles are examined, and they are as extensive in birds as they are in mammals. Some of the same neurohormones that are found in humans have been identified in nonmammals, and these neurohormones produce similar effects on cells of the pituitary as described above for mammals.

Neurosecretory centres

Two or more neurohormonal peptides with chemical and biologic properties similar to those of mammalian oxytocin and vasopressin are secreted by the vertebrate hypothalamus (except in Agnatha fishes, which produce only one). The oxytocin-like peptide is usually isotocin (most fishes) or mesotocin (amphibians, reptiles, and birds). The second peptide is arginine vasotocin, which is found in all nonmammalian vertebrates as well as in fetal mammals. Chemically, vasotocin is a hybrid of oxytocin and vasopressin, and it appears to have the biologic properties of both oxytocin (which stimulates contraction of muscles of the reproductive tract, thus playing a role in egg-laying or birth) and vasopressin (with either diuretic or antidiuretic properties). The functions of the oxytocin-like substances in nonmammals are unknown.

Tropic hormones

The pituitary glands of all vertebrates produce essentially the same tropic hormones: thyrotropin (TSH), corticotropin (ACTH), melanotropin (MSH), prolactin (PRL), growth hormone (GH), and one or two gonadotropins (usually FSH-like and LH-like hormones). The production and release of these tropic hormones are controlled by neurohormones from the hypothalamus. The cells of teleost fishes, however, are innervated directly. Thus, these fishes may rely on neurohormones as well as neurotransmitters for stimulating or inhibiting the release of tropic hormones.

Among the target organs that constitute the hypothalamic-pituitary-target organ axis are the thyroid, the adrenal glands, and the gonads. Their individual roles are discussed below.

The thyroid axis. Thyrotropin secreted by the pituitary stimulates the thyroid gland to release thyroid hormones, which help to regulate development, growth, metabolism, and reproduction. In humans, these thyroid hormones are known as triiodothyronine (T_3) and thyroxine (T_4). The evolution of the thyroid gland is traceable in the evolutionary development of invertebrates to vertebrates (Figure 5). The thyroid gland evolved from an iodide-trapping, glycoprotein-secreting gland of the protochordates (all nonvertebrate members of the phylum Chordata). The ability of many invertebrates to concentrate iodide, an important ingredient in thyroid hormones, occurs generally over the surface of the body. In protochordates, this capacity to bind iodide to a glycoprotein and produce thyroid hormones became specialized in the endostyle, a gland located in the pharyngeal region of the head.

The endostyle

From A. Gorbman and H.A. Bern, *Textbook of Comparative Endocrinology* (© 1963); by John Wiley & Sons, Inc.

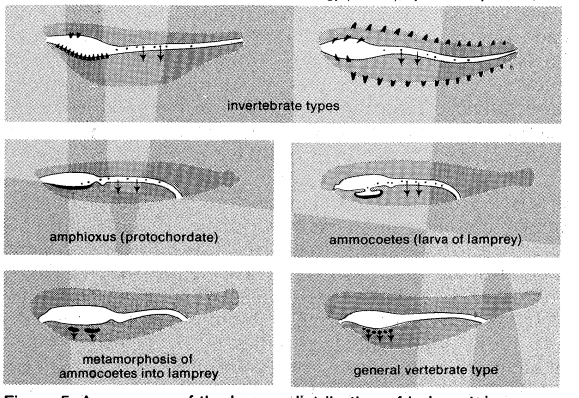


Figure 5: A summary of the known distribution of iodoproteins (shown as solid black) in the animal kingdom, suggesting a pattern of evolution of thyroid function.

When these iodinated proteins are swallowed and broken down by enzymes, the iodinated amino acids known as thyroid hormones are released. Larvae of primitive vertebrate lampreys also have an endostyle like that of the protochordates. When a lamprey larva undergoes metamorphosis into an adult lamprey, the endostyle breaks into fragments. The resulting clumps of endostyle cells differentiate into the separate follicles of the thyroid gland. Thyroid hormones actually direct metamorphosis in the larvae of lampreys, bony fishes, and amphibians. Thyroids of fishes consist of scattered follicles in the pharyngeal region. In tetrapods and a few fishes, the thyroid becomes encapsulated by a layer of connective tissue.

The adrenal axis. The adrenal axes in mammals and in nonmammals are not constructed along the same lines (Figure 6). In mammals the adrenal cortex is a separate structure that surrounds the internal adrenal medulla; the adrenal gland is located atop the kidneys. Because the cells of the adrenal cortex and adrenal medulla do not form separate structures in nonmammals as they do in mammals, they are often referred to in different terms; the cells that correspond to the adrenal cortex in mammals are called interrenal cells, and the cells that correspond to the adrenal medulla are called chromaffin cells. In primitive nonmammals the adrenal glands are sometimes called interrenal glands.

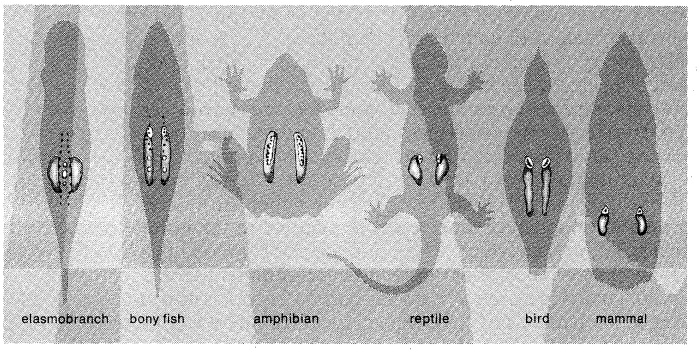


Figure 6: *Patterns of vertebrate adrenal glands.* White, interrenal or cortical tissue; black, chromaffin or medullary tissue; shaded area, kidney.

In fishes the interrenal and chromaffin cells often are embedded in the kidneys, whereas in amphibians they are distributed diffusely along the surface of the kidneys. Reptiles and birds have discrete adrenal glands, but the anatomical relationship is such that often the "cortex" and the "medulla" are not distinct units. Under the influence of pituitary adrenocorticotropin hormone, the interrenal cells produce steroids (usually corticosterone in tetrapods and cortisol in fishes) that influence sodium balance, water balance, and metabolism.

Gonadotropins

The gonadal axis. Gonadotropins secreted by the pituitary are basically LH-like and/or FSH-like in their actions on vertebrate gonads. In general, the FSH-like hormones promote development of eggs and sperm and the LH-like hormones cause ovulation and sperm release; both types of gonadotropins stimulate the secretion of the steroid hormones (androgens, estrogens, and, in some cases, progesterone) from the gonads. These steroids produce effects similar to those described for humans. For example, progesterone is essential for normal gestation in many fishes, amphibians, and reptiles in which the young develop in the reproductive tract of the mother and are delivered live. Androgens (sometimes testosterone, but often other steroids are more important) and estrogens (usually estradiol) influence male and female characteristics and behaviour.

Control of pigmentation. Melanotropin (melanocyte-stimulating hormone, or MSH) secreted by the pituitary regulates the star-shaped cells that contain large amounts of the dark pigment melanin (melanophores), especially in the skin of amphibians as well as in some fishes and reptiles. Apparently, light reflected from the surface stimulates photoreceptors, which send information to the brain and in turn to the hypothalamus. Pituitary melanotropin then causes the pigment in the melanophores to disperse and the skin to darken, sometimes quite dramatically. By releasing more or less melanotropin, an animal is able to adapt its colouring to its background.

Growth hormone and prolactin. The functions of growth hormone and prolactin secreted by the pituitary overlap considerably, although prolactin usually regulates water and salt balance, whereas growth hormone primarily influences protein metabolism and hence growth. Prolactin allows migratory fishes such as salmon to adapt from salt water to fresh water. In amphibians, prolactin has been described as a larval growth hormone, and it can also prevent metamorphosis of the larva into the adult. The water-seeking behaviour (so-called water drive) of adult amphibians often observed prior to breeding in ponds is also controlled by prolactin. The production of a protein-rich secretion by the skin of the discus fish (called "discus milk") that is used to nourish young offspring is caused by a prolactin-like hormone. Similarly, prolactin stimulates secretions from the crop sac of pigeons ("pigeon" or "crop" milk), which are fed to newly hatched young. This action is reminiscent of prolactin's actions on the mammary gland of nursing mammals. Prolactin also appears to be involved in the differentiation and function of many sex accessory structures in nonmammals, and in the stimulation of the mammalian prostate gland. For example, prolactin stimulates cloacal glands responsible for special reproductive secretions. Prolactin also influences external sexual characteristics such as nuptial pads (for clasping the female) and the height of the tail in male salamanders.

Functions of prolactin

Other vertebrate endocrine glands. *The pancreas.* The pancreas in nonmammals is an endocrine gland that secretes insulin, glucagon, and somatostatin. Pancreatic polypeptide has been identified in birds and may occur in other groups as well. Insulin lowers blood sugar (hypoglycemia) in most vertebrates, although mammalian insulin is rather ineffective in reptiles and birds. Glucagon is a hyperglycemic hormone (it increases the level of sugar in the blood).

In primitive fishes the cells responsible for secreting the pancreatic hormones are scattered within the wall of the intestine. There is a trend toward progressive clumping of cells in more evolutionarily advanced fishes, and in a few species the endocrine tissue forms only one or a few large islets. As a rule, most fishes lack a discrete pancreas, but all tetrapods have a fully formed exocrine and endocrine pancreas. The endocrine cells of all tetrapods are organized into distinct islets as described for humans, although the abundance of the different cell types often varies. For example, in reptiles and birds there is a predominance of glucagon-secreting cells and relatively few insulin-secreting cells.

Calcium-regulating hormones. Fishes have no parathyroid glands: these glands first appear in amphibians. Although the embryological origin of parathyroid glands of tetrapods is well known, their evolutionary origin is not. Parathyroid hormone raises blood calcium levels (hypercalcemia) in tetrapods. The absence in most fishes of cellular bone, which is the principal target for parathyroid hormone in tetrapods, is reflected by the absence of parathyroid glands.

Parathyroid glands

Fishes, amphibians, reptiles, and birds have paired pharyngeal ultimobranchial glands that secrete the hypocalcemic hormone calcitonin. The corpuscles of Stannius, unique glandular islets found only in the kidneys of bony fishes, secrete a peptide called hypocalcin. Fish calcitonins differ somewhat from the mammalian peptide hormone of the same name, and fish calcitonins have proved to be more potent and have a longer-lasting action in humans than human calcitonin itself. Consequently, synthetic fish calcitonin has been used to treat humans suffering from various disorders of bone, including Paget's disease (see below *The parathyroid glands: Metabolic bone disease*). The secretory cells of the ultimobranchial glands are derived from cells that migrated from the embryonic nervous system. During the development of a mammalian fetus, the ultimobranchial gland becomes incorporated into the developing thyroid gland as the "C cells" or "parafollicular cells."

Gastrointestinal hormones. Little research has been done on gastrointestinal hormones in nonmammals, but there is good evidence for a gastrinlike mechanism that controls the secretion of stomach acids. Peptides similar to cholecystokinin are also present and can stimulate contractions of the gall bladder. The gall bladders of primitive fishes contract when treated with mammalian cholecystokinin.

Other mammalian-like endocrine systems. *The renin-angiotensin system.* The renin-angiotensin system in mammals is represented in nonmammals by the juxtaglomerular cells that secrete renin associated with the kidney. The macula densa that functions as a detector of sodium levels within the kidney tubules of tetrapods, however, has not been found in fishes.

The pineal complex. In fishes, amphibians, and reptiles, the pineal complex is better developed than in mammals. The nonmammalian pineal functions as both a photoreceptor organ and an endocrine source for melatonin. Effects of light on reproduction in fishes and tetrapods are mediated at least in part through the pineal, and it has been implicated in a number of daily and seasonal biorhythmic phenomena.

Pineal functions

Prostaglandins. Many tissues of nonmammals produce prostaglandins that play important roles in reproduction similar to those discussed for humans and other mammals.

The liver. As in mammals, the liver of several nonmammalian species has been shown to produce somatomedin-like growth factors in response to stimulation by growth hormone. Similarly, there is evidence that prolactin stimulates the production of a related growth factor, which synergizes (cooperates) with prolactin on targets such as the pigeon crop sac.

Unique endocrine glands in fishes. In addition to the corpuscles of Stannius and the ultimobranchial glands, most fishes have a unique neurosecretory neurohemal organ, the urophysis, which is associated with the spinal cord at the base of the tail. Although the functions of this caudal (rear) neurosecretory system are not now understood, it is known to produce two peptides, urotensin I and urotensin II. Urotensin I is chemically related to a family of peptides that includes somatostatin; urotensin II is a member of the family of peptides that includes mammalian corticotropin-releasing hormone (CRH). There are no homologous structures to either the corpuscles of Stannius or the urophysis in amphibians, reptiles, or birds.

INVERTEBRATE ENDOCRINE SYSTEMS

Advances in the study of invertebrate endocrine systems have lagged behind those in vertebrate endocrinology, largely due to the problems associated with adapting investigative techniques that are appropriate for large vertebrate animals to small invertebrates. It also is difficult to maintain and study appropriately some invertebrates under laboratory conditions. Nevertheless, knowledge about these systems is accumulating rapidly.

Problems in invertebrate studies

All phyla in the animal kingdom that have a nervous system also possess neurosecretory neurons. The results of studies on the distribution of neurosecretory neurons and ordinary epithelial endocrine cells imply that the neurohormones were the first hormonal regulators in animals. Neurohemal organs appear first in the more advanced invertebrates (such as mollusks and annelid worms), and endocrine epithelial glands occur only in the most advanced phyla (primarily Arthropoda and Chordata). Similarly, the peptide and steroid hormones found in vertebrates are also present in the nervous and endocrine systems of many invertebrate phyla. These hormones may perform similar functions in diverse animal groups. With more emphasis being placed on research in invertebrate systems, new neuropeptides are being discovered initially in these animals, and subsequently in vertebrates.

The endocrine systems of some animal phyla have been studied in detail, but the endocrine systems of only a few species are well known. The following discussion summarizes the endocrine systems of five invertebrate phyla and the two invertebrate subphyla of the phylum Chordata, a phylum that also includes Vertebrata, a subphylum to which the backboned animals belong.

Phylum Nemertea. Nemertine worms are primitive marine animals that lack a coelom (body cavity) but differ from other acoelomates (animals that lack a coelom) by having a complete digestive tract. Three neurosecretory centres have been identified in the simple nemertine brain; one centre controls the maturation of the gonads, and all three appear to be involved in osmotic regulation.

Phylum Annelida. The cerebral ganglion (brain) of *Nereis,* a marine polychaete worm, produces a small peptide hormone called nereidine, which apparently inhibits precocious sexual development. There is a complex just beneath the brain that functions as a neurohemal organ. The epithelial cells found in this complex may be secretory as well, but this has not been proved. Neurohormones are released from the infracerebral complex into the coelomic fluid through which they travel to their targets. In the lugworm, *Arenicola,* there is evidence for a brain neuropeptide that stimulates oocyte maturation.

Phylum Mollusca. Within the phylum Mollusca, the class Gastropoda (snails, slugs) has been studied most extensively. The cerebral ganglion (brain) of several species (*e.g., Euhadra peliomorpha, Aplysia californica,* and *Lymnaea stagnalis*) secretes a neurohormone that stimulates the hermaphroditic gonad (the reproductive gland that contains both male and female characteristics); hermaphroditism is a common condition among mollusks. This gonadotropic peptide hormone (a hormone that has the gonads as its target organ) is stored in a typical neurohemal organ until its release is stimulated. For example, phototropic information detected by the so-called optic gland (located near the eye) can direct the release of the gonadotropic hormone. The gonadotropic hormones that cause egg laying in *Aplysia* and *Lymnaea* have been isolated, and they are very similar small peptides. The hermaphroditic gonad of *Euhadra* secretes testosterone (identical to the vertebrate testosterone), which stimulates formation of a gland that releases a pheromone for influencing mating behaviour. The optic gland of the octopus (of the class Cephalopoda) influences development of the reproductive organs on a seasonal basis. It is not known, however, whether any neurohormones are involved or whether this is purely a neurally controlled event.

The hermaphroditic gonad

Phylum Arthropoda. The arthropods are the largest and most advanced group of invertebrate animals, rivaling and often exceeding the evolutionary success of the vertebrates. Indeed, the arthropods are the most successful ecological competitors of humans. There are several major subdivisions, or classes, within the phylum Arthropoda, with the largest being Insecta (insects), Crustacea (crustaceans, including crabs, crayfishes, and shrimps), and Arachnida (arachnids, including the spiders, ticks, and mites). Even within these major classes, few species have been studied. Those that have been studied are large insects (*e.g.,* cockroaches, grasshoppers, and cecropia moths) and crustaceans.

The organizations of arthropod endocrine systems parallel those of the vertebrate endocrine system. That is, neurohormones are produced in the arthropod brain (analogous to the vertebrate hypothalamus) and are stored in a neurohemal organ (like the vertebrate neurohypophysis). The neurohemal organ of insects may have an endocrine portion (like the vertebrate adenohypophysis), and hormones or neurohormones released from these organs may stimulate other endocrine glands as well as nonendocrine targets. A general description of the endocrine systems of insects and crustaceans is given below.

Endocrine system organization

Class Insecta. Neurosecretory, neurohemal, and endocrine structures are all found in the insect endocrine system (Figure 7). There are several neurosecretory centres in the brain, the largest being the pars intercerebralis. The paired corpora cardiaca (singular, corpus cardiacum) and the paired corpora allata (singular, corpus allatum) are both neurohemal organs that store brain neurohormones, but each has some endocrine cells as well. The ventral nerve cord and associated ganglia also contain neurosecretory cells and have their own neurohemal organs; *i.e.,* the multiple perisympathetic organs located along the ventral nerve cord. The insect endocrine system produces neurohormones as well as hormones that control molting,

diapause, reproduction, osmoregulation, metabolism, and muscle contraction (Table 2).

Table 2: Processes in Insects Controlled by Neuropeptides of the Brain and/or Ventral Cord

Molting
Diapause
Sexual differentiation
Vitellogenesis
Sexual behaviour
Egg laying
Metabolism
Water and salt regulation
Heart rate
Rhythms in activity levels
Colour changes

Molting. A peptide neurohormone that controls molting is secreted by the pars intercerebralis and is stored in the corpora cardiaca or corpora allata (depending on the group of insects). This brain neurohormone is known as the prothoracotropic hormone (PTTH), and it stimulates the prothoracic glands (also called ecdysial or molting glands). In turn, the prothoracic glands release the steroid ecdysone, which is the actual molting hormone. Ecdysone initiates shedding of the old, hardened cuticle (exoskeleton).

Prothoracotropic hormone

In the 1940s Sir Vincent (Brian) Wigglesworth discovered that distention of the abdomen of the blood-sucking hemipteran bug *Rhodnius prolixus* following consumption of a blood meal sends neural impulses to the brain and triggers the release of PTTH. A similar mechanism has been found in a herbivorous (plant-eating) hemipteran as well. Size seems to trigger molting in lepidopterans (moths, butterflies), although the mechanism is not understood. Each molt is aided by a small amount of juvenile hormone (JH) secreted by endocrine cells of the corpora allata. Without JH during a critical time of the intermolt period of the last larval stage, either a pupa stage (diapause, or a resting state) or an adult stage is achieved. Juvenile hormone also keeps the epidermis in a larval state and causes it to secrete larval cuticle. Without JH, the epidermis changes and secretes the adult cuticle type. Three different closely related forms of JH have been isolated from seven major insect orders.

From (top) *Structure and Function in the Nervous Systems of Invertebrates* by Theodore Holmes Bullock and G. Adrian Horridge, W.H. Freeman and Company, copyright © 1965; (bottom) A. Gorbman and H.A. Bern, *Textbook of Comparative Endocrinology* (© 1963), by John Wiley & Sons, Inc.

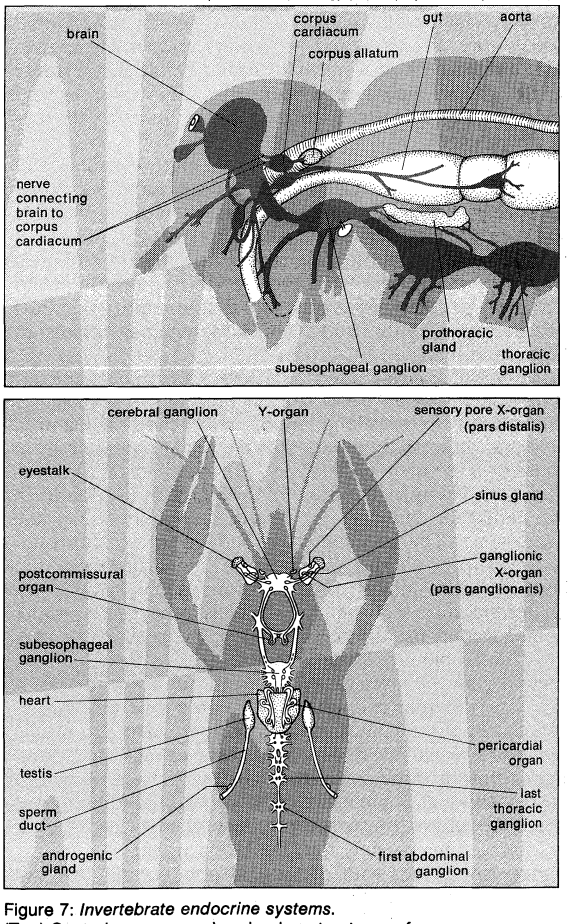


Figure 7: *Invertebrate endocrine systems.*
(Top) Central nervous and endocrine structures of a generalized insect; (bottom) endocrine system of a generalized male crustacean.

Diapause. Some insects enter diapause during development. Diapause is characterized by cessation of development or reproduction, decrease in water content (dehydration), and reduction in metabolic activities. It usually is preceded by an accumulation of nutrients resulting in hypertrophy of the fat bodies. Environmental factors (such as temperature, photoperiod, and food availability) cause storage of neurohormones, and the corpora allata become inactive. Termination of diapause can be brought about by reversing the environmental conditions that induced the diapause. Although juvenile hormone can terminate diapause, it triggers diapause in some insects. The stage of the life history may be important in determining the role of JH. For example, in imaginal diapause (characterized by cessation of reproduction in the imago, or adult), the absence of JH initiates diapause. In lepidopterans, a peptide that initiates diapause has been isolated from the subesophageal ganglion.

Effect of environmental factors

Reproduction. In some insects the pars intercerebralis secretes a neurohormone that stimulates vitellogenesis by the fat body (vitellogenesis is the synthesis of vitellogenin, a protein from which the oocyte makes the egg proteins). This neurohormone is stored in either the corpora cardiaca or the corpora allata, depending on the species. Uptake of vitellogenin by the ovary is enhanced by JH. In most insects, JH also stimulates vitellogenin synthesis by the fat body. There is evidence that other neurohormones secreted by the pars intercerebralis also influence reproduction. Some induce other tissues to secrete pheromones that influence reproductive behaviour of other individuals. In the live-bearing tsetse fly, *Glossina,* a neurohormone released from the corpora allata stimulates milk glands that provide nourishment to the developing larvae.

Osmoregulation. All insects produce a diuretic hormone and many produce an antidiuretic hormone as well. Insects feeding exclusively on a liquid diet (such as plant sap or blood) have only the diuretic hormone that allows them to eliminate excess fluid and salts through the malpighian tubules (the insect kidney). These osmoregulatory neurohormones are produced both in the brain and in the ventral nerve cord.

Myotropic and metabolic factors. One or more small peptide neurohormones are produced in the brain and ventral nervous system and are stored in the corpora cardiaca and perisympathetic organs, respectively. These myotropic factors stimulate heart rate as well as contractions of the kidney tubules and digestive tract. The corpora cardiaca were named for the heart-stimulating action produced by extracts of these organs. The glandular portion of the corpora cardiaca is thought to secrete the hyperglycemic hormone that causes a rapid increase in blood levels of trehalose, the "blood sugar" of insects. It is sometimes called the hypertrehalosemic hormone. This hypoglycemic hormone apparently is identical to the myotropic factors in at least one species, the American cockroach. An adipokinetic neurohormone released from the orthopteran corpora cardiaca (locusts, grasshoppers) causes the release of diglycerides into the blood during, and immediately after, flight. It is a peptide similar to the myotropic factors.

Class Crustacea. Among the crustaceans, the major neuroendocrine system consists of the neurosecretory X-organ and its associated neurohemal organ, the sinus gland. Both an X-organ and a sinus gland are located in each eyestalk, and together they are termed the eyestalk complex. Two endocrine glands are well known: the Y-organ and the androgenic gland (see Figure 7). As in insects, hormones and neurohormones of the crustacean regulate molting, reproduction, osmoregulation, metabolism, and

heart rate. In addition, the regulation of colour changes is well developed in crustaceans, whereas only a few insects exhibit hormonally controlled colour changes.

Molting. The steroid ecdysone secreted from the Y-organ stimulates molting. After it is released into the blood, ecdysone is converted to a 20-hydroxyecdysone, which is the active molting hormone. Secretion of ecdysone is blocked by a neurohormone called molt-inhibiting hormone, produced by the eyestalk complex. The existence of several additional molting factors has been proposed from experimental studies, and the regulation of molting may be much more complicated than suggested here.

Reproduction. The eyestalk complex appears to produce a neurohormone that inhibits vitellogenesis by the fat body and blocks vitellogenin uptake by oocytes in the ovary. Older follicles in the ovary, however, may secrete a vitellogenin-stimulating hormone that overrides the effects of the eyestalk neurohormone. In shrimps and other crustaceans that exhibit sequential hermaphroditism, the androgenic gland produces a peptide hormone that is necessary to masculinize the gonad. These animals function first as males, and later with the degeneration of the androgenic gland they become females. Surgical removal of the androgenic gland causes a precocious change of a male to a female.

Osmoregulation. There are four known sources of factors that influence water and ionic balance (osmoregulation) in crustaceans. The brain factor is known to regulate function of the antennal glands (paired "kidneys" located at the base of each antenna), the intestine, and the gills. The thoracic ganglion factor affects the stomach, intestine, and gills. Both the antennal glands and the gills are affected by a factor from the eyestalk complex. Finally, the pericardial organs (neurohemal glands located in the pericardial cavity) influence salt and water metabolism by heart muscle and gills.

Myotropic factor. Heart rate is accelerated in crustaceans by a factor released from the pericardial organs. It is not known if this factor is the same one that has osmoregulatory actions mentioned above. There is evidence to suggest that the crustacean cardioacceleratory factor is identical to one of the insect cardioacceleratory factors.

Colour changes. Several neurohormones that regulate colour changes (chromatophorotropins) by pigment cells (chromatophores) have been found in extracts of the eyestalk complex. The best known are the light-adapting hormone and the red-pigment-concentrating hormone. This latter peptide is chemically similar to the insect adipokinetic and myotropic factors. Regulation of the chromatophores allows an animal to adapt to different backgrounds by changing colours or by becoming lighter or darker.

Phylum Echinodermata. Female sea stars (starfishes) are the only echinoderms that have been studied extensively. A neuropeptide called the gonad-stimulating substance (also called the gamete-shedding substance) is released from the radial nerves into the body cavity about one hour before spawning. Gonad-stimulating substance has been reported in more than 30 species of sea star. This neuropeptide contacts the ovaries directly and causes formation of 1-methyladenine, an inducer of oocyte maturation and spawning. This same hormone has been demonstrated in the ovaries of the closely related sea urchin, where it also promotes maturation of the oocyte.

Gonad-stimulating substance

Phylum Chordata. The phylum Chordata is separated into three subgroups (or subphyla). The invertebrate subphylum Tunicata consists of the marine tunicates, including the ascidians and salps. The invertebrate subphylum Cephalochordata includes the fishlike amphioxus (or lancelet). Amphioxus is a small marine animal that closely resembles the larva of the jawless fishes (class Agnatha). The subphylum Vertebrata is the largest chordate subgroup.

Subphylum Tunicata. The ascidians (also called sea squirts) have a tadpolelike larva that lives free for a short period. The larva eventually attaches itself to a solid substrate and undergoes a marked metamorphosis into the sessile adult sea squirt. The larva and adult have a mucus-secreting gland, the endostyle, that is believed to be the evolutionary ancestor of the vertebrate thyroid gland. Metamorphosis in ascidians can be induced by application of thyroid hormones.

Neurosecretory neurons in the cerebral ganglion (brain) contain the vertebrate peptide gonadotropin-releasing hormone (GnRH). Directly adjacent to the brain is the neural (or subneural) gland that may be the forerunner of the vertebrate pituitary gland. Extracts prepared from ascidian neural glands stimulate testicular growth in toads, demonstrating the presence of a gonadotropic factor in the neural gland. A protein similar to human prolactin has been found in the neural gland of *Styela plicata.*

Subphylum Cephalochordata. The cephalochordate brain contains neurosecretory neurons that possibly are related to a structure called Hatschek's pit, located near the brain. Hatschek's pit appears to be related to the neural gland and hence to the vertebrate pituitary gland. Treatment of amphioxus with GnRH or luteinizing hormone (LH) reportedly stimulates the onset of spermatogenesis in male gonads. Furthermore, extracts prepared from Hatschek's pit can stimulate the testis of a toad. Amphioxus has a mucus-secreting endostyle like that of the ascidians. and studies have shown that the cephalochordate endostyle can synthesize thyroid hormones, too. Thus, the basic organization of the vertebrate endocrine system appears to show its early beginnings in the simple organs of these invertebrate chordates. (D.O.N.)

Hatschek's pit

The human endocrine system

GENERAL ASPECTS

Integrative functions. The endocrine systems of humans and other animals serve an essential integrative function. Inevitably, humans are beset by a variety of insults, such as trauma, infection, tumour formation, genetic defects, and emotional damage. The endocrine glands play a key role in responding to these stressful stimuli. Less obvious are the effects of subtle changes in the concentrations of key elements of the body's fluids on the storage and expenditure of energy and the steady and timely growth and development of a normal human being. These more subtle changes largely result from the monitoring by and the response, sometimes minute by minute, of the endocrine system.

The menstrual cycle in the normal, mature female and the reproductive process in males and females are under endocrine control. Beyond this, lactation and probably some forms of parental behaviour are strongly influenced by endocrine secretions. The endocrine system works in concert with the nervous and the immune systems to permit the orderly progression of human life, and these systems provide the body's bulwark against threats to health and life.

Control of endocrine system

Anatomical considerations. There are some characteristics shared by all endocrine glands. Some glands, for example the thyroid gland, are discrete, readily recognized organs with defined borders that are easily separable from adjacent structures. Others are embedded in other structures (for example, the islets of Langerhans are found in the pancreas) and may be clearly seen only under the microscope. The boundaries of endocrinology, however, have yet to be sharply defined, and endocrine tissue has been identified in surprising locations, such as the heart. Under the microscope, endocrine cells appear to be rather homogeneous, usually cuboidal in shape, with a rich supply of small blood vessels. Sometimes, as is the case in the thyroid gland, endocrine cells are intermixed with other, distinctly different endocrine cells with a different embryological origin and an entirely different set of hormonal secretions. Finally, all nerve cells are capable of secreting neurotransmitters into the synapses between adjacent nerves, although some nerve cells, for example those of the neurohypophysis (posterior pituitary gland), also secrete neurohormones directly into the bloodstream.

Endocrine glands with mixed cell populations have not evolved by chance. The hormonal secretions of one set may modulate directly the activity of adjacent cells with different characteristics. This direct action on contiguous cells of different types, which diffuses the hormone to tar-

Paracrine function

get cells without moving it through the general circulation, is known as paracrine function. Even in homogeneous glandular tissues (*i.e.,* tissues comprising one cell type), the direct proximity of the cells in some way enhances the amount of hormonal secretions since isolated cells are less vigorous in their activity, under laboratory conditions, than are sheets of attached cells, a phenomenon known as autocrine function. On the other hand, hormonal secretions themselves inhibit further hormonal secretion if they remain in the vicinity of the parent cell.

When viewed under the ultramicroscope (a microscope of extraordinary magnifying power), the endocrine cell has the fine structure that is illustrated in Figure 8. Many of the various intracellular structures, called organelles, are involved in the sequence of events that occurs during the synthesis and secretion of hormones.

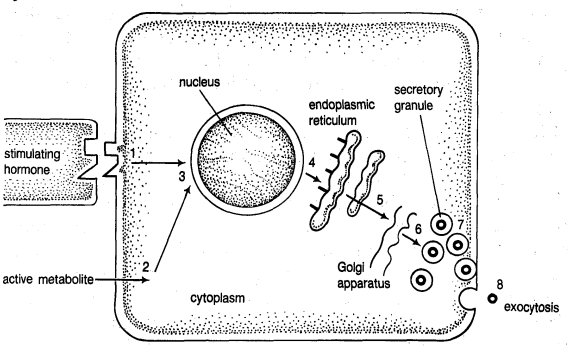


Figure 8: Intracellular structure of a typical endocrine cell.

Hormone synthesis. In the case of protein hormone synthesis, the target cell is stimulated at the cell surface (1 and 2) either by contact of a surface receptor with a stimulating hormone or by the entrance of a stimulating metabolite, such as glucose entering an insulin-producing beta cell. (It is important to note that there are also hormones and metabolites that lead to the inhibition of cell activity, but these are not discussed here.) Stimulation of the receptor at the cell surface is followed by a series of complex events within the cell membrane itself as well as within the cytoplasm of the cell. These events lead to the stimulation of DNA within the nucleus of the cell (3) to synthesize mRNA (messenger ribonucleic acid), which directs the synthesis of protein or steroid hormones.

Synthesis of mRNA

The mRNA unit contains the genetic code for the new protein. When mRNA leaves the nucleus and associates with the endoplasmic reticulum (4) in the cytoplasm, it directs the synthesis of a relatively inert precursor to the hormone, called a prohormone, from free amino acids available within the cytoplasm. The prohormone is sent to another cell organelle (5), the Golgi apparatus, where it is packaged into vesicles known as secretory granules (6). The granules migrate to the cell surface (7), and through a process known as exocytosis (8) the active hormone splits from the prohormone and is discharged through the cell wall.

In the case of steroid hormones, the precursor of all steroid hormones, cholesterol, is stored in vesicles within the cytoplasm. Through the actions of enzymes at various steps along the synthetic pathway, cholesterol is broken down and converted into steroid hormones.

Cholesterol is converted into pregnenolone in the first step of steroid biosynthesis. This action is the result of the cleaving enzyme that has been stimulated into action by corticotropin (ACTH) or angiotensin (which stimulate the adrenals), or the gonadotropins (hormones, such as LH and FSH, that stimulate the gonads). Pregnenolone is transported out of the mitochondria (where this initial step took place) to the endoplasmic reticulum (4), where it undergoes further enzymatic degradation to progesterone. At this point depending on the tissues in which the synthesis took place, progesterone is converted to the sex hormones (androgens and estrogens) or to the corticoids, mineralocorticoids, or adrenal androgens (steroid hormones of the adrenal cortex).

Regulatory mechanisms. Hormonal levels in the circulating body fluids vary in response to stimulatory or inhibitory influences acting on the hormone-producing cell. In the normal individual examined in a resting state, all circulating hormonal levels will be found to lie within a narrow normal range. Constant monitoring of hormonal supply to the tissues is essential to health, since sustained, inappropriate elevations or depressions of these levels will lead, in most instances, to disease states. Furthermore, since hormones are constantly being inactivated by tissue enzymes, the supply of hormones must be replenished regularly within the cell by synthesis and secretion.

Control of hormonal levels

The control of hormonal levels is maintained by a number of feedback devices. For target organs such as the thyroid, adrenal glands, and gonads, which also serve as endocrine glands, the hypothalamic-pituitary-target organ axis serves admirably. Other more direct feedback mechanisms, however, also operate (see Figure 2), for example, the stimulatory effect of low serum calcium levels on parathyroid glands and the stimulatory effect of elevated blood glucose levels on the beta cells of the islets of Langerhans. In another method of hormonal regulation, the metabolism of hormones after their secretion may either intensify or decrease hormonal action; for example, thyroxine may be converted in a number of tissues to triiodothyronine (T_3), a change that enhances hormonal potency by 2½ times. Alternatively, T_4 may be converted to an inactive isomer (a molecule with the same atoms but with small, biologically important differences in structure) of T_3 (reverse T_3). Finally, local effects may significantly modulate endocrine cell activity. For most endocrine cells, if the secreted hormone remains in the immediate vicinity of the cell, further synthesis of the hormone is strongly inhibited. This effect supplements the autocrine and paracrine effects on adjacent cells of a tissue.

Modes of transport. Most hormones are secreted into the general circulation to exert their effects on appropriate distant target tissues. There are important exceptions, however, in the case of self-contained portal circulations in which blood is directed to specific areas. A portal circulation begins in a capillary bed, forms into a set of veins, and then is dispersed into a second capillary bed. Thus, blood collected from the first capillary bed is directed solely into the tissues nourished by the second capillary bed.

Two portal systems

Two such portal circulations are present in the human body. One system, the hypothalamic-hypophyseal portal circulation, collects blood from capillaries originating in the hypothalamus and, through a plexus of veins surrounding the pituitary stalk, directs the blood into the substance of the anterior pituitary. In this instance, hormones secreted by the neuroendocrine cells of the hypothalamus are transported directly, via this circulatory system, to modulate the activity of the endocrine cells of the anterior pituitary. These hormones are largely, but not entirely, excluded from the general circulation. In a second system, the hepatic portal system, capillaries originating in the gastrointestinal tract and the spleen are transported through veins by way of the hepatic vein into the liver and again dispersed into hepatic capillaries. In this way hormones from the pancreatic islets of Langerhans, such as insulin and glucagon, are directed into the substance of the liver in high concentration before being distributed through the general circulation.

Carrier proteins

A further refinement in hormone transport is provided by circulating carrier proteins. These substances, manufactured and secreted by the liver, provide sites to which steroid and thyroid hormones are bound. Carrier proteins include the binding globulins that bind sex hormones from the gonads, and transcortin, to which hormones from the adrenal cortex are bound. In addition, there are two sets of proteins, the prealbumins and the thyroxine-binding globulins, which transport the thyroid hormones, T_4 and T_3. Furthermore, there is evidence that other protein hormones, such as growth hormone, also are bound to specific transport proteins. Indeed, it is the rule that important biologic substances are bound to specific carrier proteins as they course through the circulation.

Protein-bound hormones are in equilibrium with a much smaller concentration of free circulating hormones. As a free hormone leaves the circulation to exert its action on a tissue, an equal amount of hormone is immediately freed

from its binding protein. Thus, the carrier proteins serve as a depot within the bloodstream to maintain a normal concentration of the biologically important free hormone. A final refinement of this system is that the concentration of carrier proteins is similarly hormone-dependent. Estrogens are known to increase the secretion and concentration of essentially all carrier proteins, while androgens generally have an opposite effect.

The affinity of hormones for these binding proteins is not constant. The thyroid hormone T_4, for example, is far more tightly bound than is the hormone T_3, with the result that T_3 is more readily released as a free molecule and has easier access to tissues. Similarly, some drugs, such as phenytoin (Dilantin), have a molecular configuration that permits them to compete with thyroxine for binding to thyroxine-binding globulin. When present in high concentrations, such a drug may successfully displace the level of bound hormones in the blood.

Biorhythms. Some hormones, for example insulin, are secreted in brief spurts every few minutes. Presumably, the time between spurts is a reflection of the lag time necessary for the insulin-secreting cell to sense a change in blood sugar concentration. Other hormones, particularly those of the pituitary, are secreted in pulses that may occur at roughly hourly intervals. Apparently, pulsated secretion is a necessary requirement for pituitary gonadotropin secretion. When stimulated at about 80-minute intervals by the injection of a hypothalamic gonadotropin-releasing hormone (GnRH), pituitary gonadotropin secretion increases incrementally to high levels. If, however, gonadotrophs are subjected to a continuous, nonpulsatile injection of GnRH, gonadotropin secretion is completely inhibited.

Circadian rhythms

Superimposed on these pulsatile secretions are, for many hormones, changes in hormonal levels that occur at roughly 24-hour intervals. These periodic changes are called circadian rhythms. An example is shown in Figure 9, which illustrates the circadian changes in the blood concentration of cortisol, the major steroid hormone secreted by the adrenal cortex. Low levels occur during sleep, with a rapid rise in the early morning hours, followed by a graduated descent during the day, with intermediate elevations during meal times. This particular rhythm is dependent on night-day cycles and persists for some days after jet-plane travel into different time zones. The transitional period is reflected in the well-known phenomenon of jet lag. Other hormones follow other circadian rhythms. Pituitary growth hormone, prolactin, and the gonadotropins rise to their highest diurnal (daily) levels shortly after the onset of sleep and, in the case of gonadotropins, this sleep-related elevation is the first biochemical sign of the onset of puberty.

From E.D. Weitzman *et al.*, "Twenty-four-hour Patterns of the Episodic Secretion of Cortisol in Normal Subjects," *Journal of Clinical Endocrinology and Metabolism*, vol. 33, pp. 13–22, © by The Endocrine Society, 1971

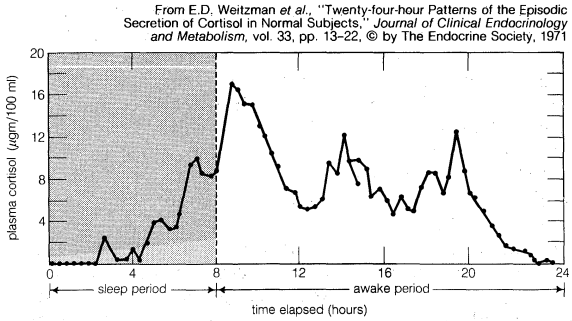


Figure 9: Circadian rhythm, a graphic depiction of cortisol values over a 24-hour period.

Monthly biorhythms are reflected in women in the menstrual cycle. Less obvious are seasonal cycles that occur in the secretion of thyroid hormones and testosterone. Finally, puberty itself is a complex, timed phenomenon that is thought to be associated with a reduction in secretion of melatonin, a hormone secreted by the pineal gland. This gland serves a regulatory function for many of the biorhythms, particularly those related to gonadal function.

ENDOCRINE DYSFUNCTION

Endocrine hypofunction and receptor defects. There are occasions when the body is best served by reducing the amount of hormone secreted by an endocrine gland (hypofunction). For example, the secretion of thyroid hormones decreases with protracted fasting. Because the thyroid hormones control energy expenditure, there is survival value in slowing the body's metabolism when there is no intake of food. Thus, there is a distinction between compensatory endocrine hypofunction and true endocrine dysfunction. Only those forms of hypofunction that reflect disease states are discussed in general terms below. Detailed descriptions of specific endocrine deficiency states are given in later sections devoted to each of the individual endocrine organs.

Destruction of endocrine cells

Endocrine glands may be destroyed in a variety of ways, but complete destruction is difficult. For most endocrine organs, at least 90 percent of the gland must be destroyed before a significant illness occurs. In the case of paired endocrine organs (parathyroids, adrenal glands, and the gonads) the removal of one of the pair is followed by a prompt compensatory increase in the activity of the remaining gland, so that an affected individual continues in good health.

Physical trauma (including surgical trauma and severe hemorrhage within the gland substance) may destroy any endocrine gland. Similarly, an invasion, known as infiltration, by cancer cells, inflammatory cells, large amounts of metal such as iron or copper, or by an abnormal protein, such as amyloid, may also seriously impair endocrine function. Bacterial infections (such as tuberculosis) and viral infections (such as mumps) also may lead to endocrine deficiencies. Although radiation damage from either X rays or radioactive elements is a well-recognized cause of hormonal deficiencies, both avenues have been adapted as forms of treatment when the problem is endocrine hyperfunction (excessive secretion by an endocrine gland).

Autoimmunity

Last, and perhaps most important, there is a growing understanding of an extraordinary phenomenon, known as autoimmunity, as a cause of endocrine deficiency. Certain antibodies generated by the body against its own tissues (see IMMUNITY), have been found to be active against certain endocrine tissues. Thus, not only are specific antibodies formed against specific endocrine glands, but there are also antibodies that affect specific aspects of endocrine function. For instance, in the case of the thyroid there are cytotoxic antibodies that eventually destroy the gland by attacking the cells; there are blocking antibodies that can, in effect, inactivate thyroid cell surface receptors and cause hypothyroidism; and there are stimulatory antibodies, which are a major cause of hyperthyroidism.

Constant exposure of an endocrine gland to blocking antibodies results in a reduction in its cell size and number, a condition known as atrophy. If long-lasting, atrophy may lead to irreversible destruction of the gland. Another cause of atrophy is a receptor defect that results when autoantibodies exert their actions against endocrine surface cell receptors. This kind of receptor damage has been found in females with premature ovarian failure associated with menopause, which can occur as early as the teenage years. It remains debatable whether a natural menopause is an example of hypofunction of the ovary, which should be viewed as pathological, or whether it represents another example of compensatory hypofunction with "survival value."

Endocrine atrophy is also associated with a number of forms of developmental failure, such as chromosomal abnormalities. For example, in Turner's syndrome, the Y chromosome of the two sex chromosomes, X and Y, is missing, resulting in the body configuration and orientation of a female despite the absence of functioning ovarian tissue. Another example is Klinefelter's syndrome, in which an extra X chromosome is added to the normal male complement of an X and a Y chromosome, leading to the development of an individual who appears as a feminized male and has some features of male hormone deficiency.

Secondary endocrine hypofunction

Secondary endocrine hypofunction is another distinct category of endocrine dysfunction, in which the gland is basically intact but lies dormant because it is either not stimulated or is directly inhibited. An important characteristic of this form of deficiency is that it is reversible, re-

turning to normal with the removal of the inhibiting agent. Secondary endocrine hypofunction results, for example, from the loss of a stimulating (tropic) hormone when the pituitary gland is completely destroyed. The loss of thyrotropin, corticotropin, and gonadotropins leads to hypofunction of the thyroid, adrenal, and gonads. Endocrine hypofunction may also occur as a result of exposure to excessive amounts of a hormone. In a patient taking large amounts of thyroid hormone the secretion of the thyroid-stimulating hormone thyrotropin by the anterior pituitary gland (see above *The nature of endocrine regulation*) will be inhibited, a change that puts the thyroid gland to rest.

Changes in the biochemical environment of the thyroid gland may also lead to a reduction in function. A well-known example is that of hypothyroidism due to iodine deficiency. Since iodine is an integral part of the thyroid hormone molecule, iodine deficiency leading to cretinism is common in those areas of the world in which salt contains little or no iodine. Drugs may also lead to a functional endocrine deficiency; such is the case in patients with manic-depressive psychosis treated with lithium, a drug that blocks thyroid gland activity. Finally, an excess of one hormone may lead to a deficiency of another. For example, overproduction of a pituitary hormone, prolactin, results in a secondary suppression of gonadal function, leading to amenorrhea in females and impotence in males. These changes are readily reversed when the level of prolactin in the bloodstream is returned to normal.

Iodine deficiency

Hormonal deficiency states can also occur from defective hormonal action on target organs. This concept was first proposed by an American clinical endocrinologist, Fuller Albright, and his associates in 1942. They studied a young woman who manifested all of the signs of deficiency of parathyroid hormone (PTH) but who, unlike the usual such patient, did not show any improvement after the injection of an extract of parathyroid gland. Albright termed this variant pseudohypoparathyroidism and postulated that "the disturbance is not a lack of PTH but an inability to respond to it." Direct evidence supporting this suggestion emerged only decades later, and other examples of unresponsiveness of target tissues to hormones have been documented. Thus, for example, an absence of receptors for male hormones makes individuals who are in genetic terms outwardly male appear to be female. Some diabetics do not respond to large quantities of insulin because they lack effective receptors in target cells for binding insulin. More common is resistance to insulin in diabetics due to the appearance of anti-insulin antibodies following insulin injections. Other antihormone antibodies may appear spontaneously and provoke endocrine deficiencies. Finally, there are rare instances in which hormone synthesis is abnormal so that the hormone secreted is chemically defective enough to impair its action on target tissues.

It should be noted that endocrine deficiencies may result from transmission of harmful materials from a mother to her fetus by way of the placenta. Toxic agents include autoantibodies and drugs, both of which cross the placenta readily and may damage the fetus even though, on occasion, the mother may remain unaffected.

The aging process

Because in many countries larger proportions of the populations are aging, an intensive search for the causes of aging processes has been instituted. An early popular theory was that aging resulted from multiple endocrine deficiencies. This idea has been discarded by most investigators. The only documented endocrine failure associated with age is the loss of ovarian hormones at the time of, and subsequent to, the menopause. Even here, however, the ovary continues to produce reduced amounts of estrogens. In general, endocrine function is highly variable in the aged. For most glands there is either no change or a modest reduction in endocrine secretions, but in the case of the pituitary gonadotropins, progressive gonad failure associated with aging results in pituitary hypersecretion (excessive secretion). Whether the changes observed have survival value is not known.

Endocrine hyperfunction. With excessive stimulation from any of a variety of causes, endocrine glands may become overactive, resulting in hypertrophy (increase in size of each cell) and hyperplasia (increase in cell numbers). The result is that the gland becomes enlarged. With continued stimulation, some undefined barrier is breached, and the hyperplastic glands undergo a transformation and begin uncontrolled multiplication of abnormal cells, termed neoplastic (tumorous). Because endocrine neoplasms are largely autonomous, they are far less sensitive to any inhibition of their hormonal secretions through negative feedback control. The result is that benign endocrine neoplasms (adenomas) persistently secrete excess hormone. Continued hyperstimulation causes some adenomas to undergo an additional change to a truly malignant neoplasm (a carcinoma), which is not only hyperfunctional but also is capable of invading adjacent structures and metastasizing (transferring) to distant organs, with the threat of causing death. Sometimes tumours of several endocrine glands occur simultaneously (see below *Ectopic hormone and polyglandular disorders*), which has been described as a syndrome (constellation of symptoms) called hereditary multiple endocrine neoplasia. It should also be noted, however, that many endocrine neoplasms produce no hormones whatsoever.

Neoplasms

Excess hormone secretion and the resultant symptoms may be produced by endocrine hyperplasia alone. One example of this occurs when a circulating autoantibody binds to receptors in the thyroid gland and causes the hypersecreting, hyperplastic thyroid typical of Graves' disease. Other syndromes of endocrine hyperfunction may result when a small endocrine tumour, innocuous in itself, secretes excessive amounts of a stimulatory hormone, which then provokes a secondary hyperplasia of a target gland. The classic instance is Cushing's disease, in which a small pituitary tumour produces excess quantities of adrenocorticotropin (ACTH) and leads to hyperplasia of both adrenal glands. The result is oversecretion of the hormones of the adrenal cortex, with striking consequences. A rare example of adenoma formation resulting from unremitting stimulation is found in patients with long-standing thyroid hormone deficiency. Through negative feedback mechanisms, the pituitary cells that secrete thyroid-stimulating hormone (TSH) become hyperplastic and eventually are transformed into TSH-producing adenomas of the pituitary.

Cushing's disease

Some endocrine tumours not only produce excess quantities of the expected hormone but also excess amounts of a hormone that is normally secreted by an entirely different endocrine gland. Thus, a medullary carcinoma of the thyroid originates from cells that normally produce calcitonin, a hormone which acts to lower the concentration of calcium in the blood. This tumour may hypersecrete not only calcitonin but also ACTH, normally a secretory product of cells of the pituitary gland. In addition, tumours arising from tissues that ordinarily have no endocrine function may secrete one or more hormones. A typical example is that of a cancer of the lung, which may produce one or more of an array of hormones, most commonly antidiuretic hormone. Such neoplasms are called ectopic (displaced) hormone-producing tumours.

The source of stimulation of hyperplastic glands is often known, as in the case of the parathyroid hyperplasia that follows persistently low levels of serum calcium in patients with severe kidney disease. In other instances, however, no cause has been identified; in these cases, the cause is said to be idiopathic. An example of idiopathic hyperplasia is the increase in the number of insulin-producing beta cells in the islets of Langerhans, which produces severe brain-damaging hypoglycemia (lowering of the blood sugar) in infants.

Some hormones exert their regulatory actions through an agonist/antagonist relationship to tropic hormones and receptor sites of the target cell. An agonist is a substance (for example, a hormone or a drug) that binds with specific receptors on target cells and elicits a response. An antagonist (also a hormone or drug) is a substance with a molecular structure similar enough to that of the agonist to compete with it and bind to the same specific receptors, although, once bound, it does not elicit a response. The actions of the antagonist hormone may modify the hypersecretion of the agonist hormone by binding with some of the available receptor sites, and the loss of an antagonist may

Agonists and antagonists

lead to effective hyperfunction of the agonist. An example is a person who has a deficiency of the adrenal cortex, which produces hormones that are sharply antagonistic to the action of insulin. When fasting or when injected with only a small amount of insulin, such individuals suffer the effects of severe hypoglycemia.

The general aspects of the human endocrine system discussed above may now be applied specifically in the more detailed discussion of individual endocrine glands.

THE HYPOTHALAMUS

Anatomy. The hypothalamus is an integral part of the substance of the brain. A small cone-shaped structure, it projects downward, ending in the pituitary (infundibular) stalk, a tubular connection to the pituitary gland. Figure 10 shows the relationship of these small structures in a lateral midline projection of the human head using the technique of magnetic resonance imaging. The round bony cavity containing the pituitary gland is called the sella turcica. The posterior portion of the hypothalamus, called the median eminence, contains many neurosecretory cells. Important adjacent structures include the mamillary bodies, the third ventricle, and the optic chiasm, the last being of particular concern to physicians because pressure from expanding tumours or inflammations in the hypothalamus or pituitary gland may result in severe visual defects or total blindness. Above the hypothalamus is the thalamus. (For discussion of the function of these surrounding structures, see NERVES AND NERVOUS SYSTEMS.)

Location

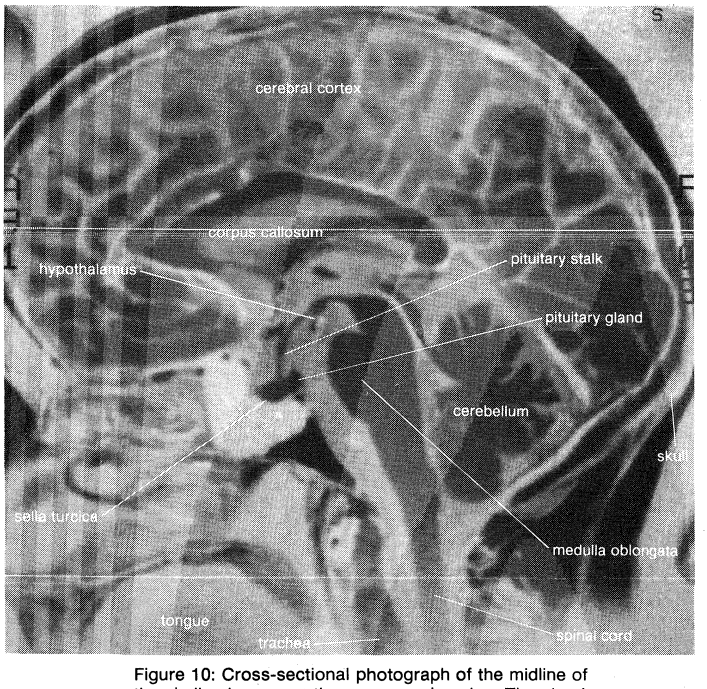


Figure 10: Cross-sectional photograph of the midline of the skull using magnetic resonance imaging. The glands within the brain and their physical interrelationships are illustrated (see text).

Regulation of hormone secretion. The hypothalamus, like the rest of the brain, consists of interconnecting nerve cells (neurons) with a rich blood supply. To understand hypothalamic function it is necessary to define the various forms of neurosecretion. First, there is neurotransmission, which occurs throughout the brain and is the process by which one nerve cell communicates with another at an intimate intermingling of projections from the two cells (a synapse). This transmission of an electrical impulse from one cell to another requires the secretion of a chemical substance from a long projection from one nerve cell body (the axon) into the synaptic space. The chemical substance that is secreted is called a neurotransmitter. The process of synthesis and secretion of neurotransmitters is similar to that shown in Figure 8 with the exception that neurosecretory granules migrate through lengths of the axon before being discharged into the synaptic space.

Classical neurotransmitters

Neurologists have long been aware of four classical neurotransmitters: epinephrine, norepinephrine, serotonin, and acetylcholine, but recently there have emerged a large number of additional neurotransmitters, of which an important group is the neuropeptides. While bioamines and neuropeptides function as neurotransmitters, some of them also perform the role of neuromodulators; they do not act directly as neurotransmitters but rather as inhibitors or stimulators of neurotransmission. Well-known examples are the opioids (for example, enkephalins), so named because they are the naturally occurring peptides with a strong affinity to the receptors that bind opiate drugs such as morphine and heroin. In effect, they are the body's opiates.

Thus the brain, and indeed the entire central nervous system, consists of an extraordinary network of neurons interconnected by neurotransmitters. The secretion of specific neurotransmitters, modified by neuromodulators, lends an organized, directed function to the overall system. These neural connections extend upward from the hypothalamus into other key areas, including the cerebral cortex. The result is that intellectual and functional activities as well as external influences, including stresses, can be funneled into the hypothalamus and thence to the endocrine system, from which they may exert effects on the body.

Neurohormones

In addition to secreting neurotransmitters and neuromodulators, the hypothalamus synthesizes and secretes a number of neurohormones. The neurons secreting neurohormones are true endocrine (neurohemal) cells in the classical sense since secretory granules containing neurohormones travel from the cell body through the axon to be extruded, where they enter directly a capillary network, thence to be transported through the hypophyseal-portal circulation to the anterior pituitary gland.

Finally, the neurohypophysis, or posterior lobe of the pituitary gland, provides the classical example of neurohormonal activity. The secretory products, mainly vasopressin and oxytocin, are extruded into a capillary network, which feeds directly into the general circulation.

The existence of hormones of the hypothalamus was predicted well before they were detected and chemically characterized, and they have been studied intensively by many investigators. Two groups of American investigators, led by Andrew Schally and Roger Guillemin, respectively, shared the Nobel Prize for Physiology or Medicine for 1977 for their research on pituitary hormones.

Releasing hormones

These neurohormones are known as releasing hormones because the major function generally is to stimulate the secretion of hormones originating in the anterior pituitary gland. They consist of simple peptides (chains of amino acids) ranging in size from only three amino acids (thyrotropin-releasing hormone) to 44 amino acids (growth hormone-releasing hormone).

Hormones. *Thyrotropin-releasing hormone.* Thyrotropin-releasing hormone (TRH), a neurohormone, is the simplest of the hypothalamic neuropeptides. It consists essentially of three amino acids in the sequence glutamic acid–histidine–proline. The simplicity of this structure is deceiving for TRH is involved in an extraordinary array of functions. Not only is it a neurohormone that stimulates the secretion of thyroid-stimulating hormone from the pituitary, and, quite independently, the secretion of another pituitary hormone called prolactin, but it also subserves other central nervous system activities because it is a widespread neurotransmitter or neuromodulator within the brain and spinal cord. There is evidence that TRH is involved in the control of body temperature and that it has psychological and behavioral effects, at least in animals. It may also have therapeutic value. There are studies suggesting that it mitigates the damage induced by spinal cord injury and that it leads to some improvement in the nervous disease known as amyotrophic lateral sclerosis (Lou Gehrig's disease).

These multiple effects are less surprising when it is considered that TRH appeared very early in the evolutionary scale of vertebrates and that, while the concentration of TRH is greatest in the hypothalamus, the total amount of TRH in the remainder of the brain far exceeds that of

the hypothalamus. The TRH-secreting cells are subject to stimulatory and inhibitory influences from higher centres in the brain and they also are inhibited by circulating thyroid hormone. In this way TRH forms the topmost segment of the hypothalamic-pituitary-thyroid axis.

Gonadotropin-releasing hormone. Gonadotropin-releasing hormone (GnRH), a neurohormone also known as luteinizing hormone-releasing hormone (LHRH), is a peptide chain of 10 amino acids. It stimulates the synthesis and release of the two pituitary gonadotropins, luteinizing hormone (LH) and follicle-stimulating hormone (FSH). While some investigators hold that there are two types of GnRH, most evidence supports the conclusion that only one type of neuropeptide stimulates the release of the two gonadotropins and that the variations in levels of both gonadotropins in the circulation are due to other modulating factors.

GnRH effects

Characteristic of all releasing hormones and most striking in the case of GnRH is the phenomenon of pulsatile secretion. In normal individuals, GnRH is released in spurts at intervals of about 80 minutes. The pulsatile administration of GnRH in large doses results in an ever-increasing concentration of gonadotropins in the blood. In striking contrast, the constant infusion of GnRH suppresses gonadotropin secretion. This phenomenon is advantageous for persons for whom suppression is important. Such persons include children with precocious puberty, and elderly men with cancer of the prostate. On the other hand, pulsatile administration of GnRH is efficacious in men or women in whom deficiency of gonadal function is due to impaired secretion of GnRH. A potential application of this phenomenon is the synthetic modifications of GnRH as a male as well as a female contraceptive.

Limbic system

Neurons that secrete GnRH have connections to an area of the brain known as the limbic system, which is heavily involved in the control of emotions and sexual activity. Studies in rats deprived of pituitary glands and ovaries but maintained on physiological amounts of female hormone (estrogen) have demonstrated that the injection of GnRH results in complex and striking changes in posture characteristic of the receptive female stance for sexual intercourse.

Some individuals have hypogonadism (in which the functional activity of the gonads is decreased and sexual development is inhibited) due to a congenital deficiency of GnRH, which responds to treatment with GnRH. Most of these people also suffer from hypothalamic disease and are deficient in other releasing hormones as well, but there are also individuals in whom GnRH deficiency is isolated and associated with a loss of the sense of smell (anosmia). Abnormalities in the pulses of GnRH secretion result in subnormal fertility, abnormal or absent menstruation, and possibly cystic disease of the ovary or even ovarian cancer.

Corticotropin-releasing hormone. Corticotropin-releasing hormone (CRH), a neurohormone, is a large peptide consisting of a single chain of 41 amino acids. It stimulates not only secretion of corticotropin in the pituitary gland but also the synthesis of corticotropin in the corticotropin-producing cells (corticotrophs) of the anterior lobe of the pituitary gland. Many factors, both neurogenic and hormonal, regulate the secretion of CRH, since CRH is the final common element directing the body's response to all forms of stress, whether physical or emotional, external or internal. (This key role of CRH in the hypothalamic-pituitary-adrenal axis is discussed below in connection with the adrenal gland.) Among the hormones that play an important role in modulating the influence of CRH is cortisol, the major hormone secreted by the adrenal cortex, which, as part of the negative feedback servomechanism (exerting control over another system through negative feedback), blocks the secretion of CRH. Vasopressin, the major regulator of the body's excretion of water, has an additional ancillary role in stimulating the secretion of CRH.

Cortisol

Excessive secretion of CRH leads to an increase in the size and number of corticotrophs in the pituitary gland, often resulting in a pituitary tumour. This, in turn, leads to excessive stimulation of the adrenal cortex, resulting in high circulating levels of adrenocortical hormones, the clinical manifestations of which are known as Cushing's syndrome. Conversely, a deficiency of CRH-producing cells can, by a lack of stimulation of the pituitary and adrenal cortical cells, result in adrenocortical deficiency. (These conditions are discussed below.)

Growth hormone-releasing hormone. Like CRH, growth hormone-releasing hormone (GHRH) is a large peptide. A number of forms have been described that differ from one another only in minor detail and in the number of amino acids (varying from 37 to 44). Unlike the other neurohormones, GHRH is not widely distributed in other parts of the brain. It is stimulated by stresses, including physical exercise, and secretion is blocked by a powerful inhibitor called somatostatin (see below *Somatostatin*). Negative feedback control of GHRH secretion is mediated largely through compounds called somatomedins, growth-promoting hormones that are generated when tissues are exposed to growth hormone itself.

Acromegaly

An excess of circulating growth hormone in adults leads to a condition called acromegaly. Rarely, a benign tumour, called a hamartoma, located in the hypothalamus may produce excessive amounts of GHRH, leading to acromegaly. Equally rare are tumours arising in the islets of Langerhans of the pancreas that may secrete excessive quantities of GHRH. Indeed, GHRH was first successfully isolated and analyzed from such an ectopic (abnormally positioned) hormone-producing tumour. Isolated deficiency of GHRH (in which there is normal functioning of the hypothalamus except for this deficiency) may be the cause of one form of dwarfism, a general term applied to all individuals with abnormally small stature.

Somatostatin. Somatostatin refers to a number of polypeptides consisting of chains of 14 to 28 amino acids. The name was coined when its discoverers found that an extract of the hypothalamus strongly inhibited the release of growth hormone from the pituitary gland. Somatostatin is also a powerful inhibitor of pituitary TSH secretion. Somatostatin, like TRH, is widely distributed in the central nervous system and in other tissues. It serves an important paracrine function in the islets of Langerhans, by blocking the secretion of both insulin and glucagon from adjacent cells. Somatostatin has emerged not only as a powerful blocker of the secretion of GH, insulin, glucagon, and other hormones but also as a potent inhibitor of many functions of the gastrointestinal tract, including the secretion of stomach acid, the secretion of pancreatic enzymes, and the process of intestinal absorption. Despite these multiple, widespread actions, the term somatostatin has persisted because of its major role as a regulator of GH secretion, and impaired somatostatin secretion may cause some forms of hypersecretion of growth hormone.

No examples of somatostatin deficiency have been found, but a tumour called a somatostatinoma has been well characterized in a number of patients. Persons with a somatostatinoma have cramping abdominal pain, persistent diarrhea, a mild elevation of blood glucose levels, and sudden flushing of the skin.

Prolactin-inhibiting and -releasing hormones. The hypothalamic regulation of prolactin secretion from the pituitary is different from the hypothalamic regulation of other pituitary hormones in two respects. First, the hypothalamus primarily inhibits rather than stimulates the release of prolactin from the pituitary (the hypothalamus stimulates the release of all other pituitary hormones). Thus, if pituitary cells are removed from the influence of the hypothalamus, few or none of the pituitary hormones are secreted, except for prolactin, which continues to be secreted by the prolactin-secreting cells (lactotrophs). Second, this major inhibiting factor is not a neuropeptide, but rather the neurotransmitter dopamine, a fact exploited in afflicted persons by physicians who are able to reduce abnormally high concentrations of prolactin by using drugs that mimic the prolactin-inhibiting effects of dopamine. Another prolactin-inhibiting factor (PRF) comes into play primarily during pregnancy and lactation. Prolactin-stimulating factors also exist, among them TRH.

Regulation of prolactin secretion

Prolactin deficiency is known to occur, but only rarely. Excessive prolactin production (hyperprolactinemia) is a common endocrine abnormality, and the prolactinoma is the most frequently encountered pituitary tumour.

Gastrointestinal neuropeptides. Although modern endocrinology began with the discovery that a substance, secretin, secreted into the blood from the cells lining the gastrointestinal tract stimulates the secretion of pancreatic juices, little attention was subsequently paid to gastrointestinal hormones. When investigators began to examine the distribution of neuropeptides within the body, however, there emerged a bewildering variety of these hormones, not only within the brain but also in the lining of the gastrointestinal tract and in other organs. The list includes glucagon, the enkephalins, secretin, cholecystokinin, gastrin, calcitonin, angiotensin, substance P, pancreatic polypeptide, neuropeptide Y (a human variant of a peptide called bombesin), delta-sleep-inducing peptide, and vasoactive intestinal peptide. The actions and interactions of these hormones both in the intestinal tract and in the brain are complex and are the subject of continuing investigations. Briefly, these peptides play important roles in the transmission and inhibition of pain stimuli, in eating and drinking behaviour, in memory and learning, in the regulation of body temperature, in the induction of sleep, and in sexual behaviour.

THE ANTERIOR PITUITARY

Anatomy. The pituitary gland lies at the base of the skull, nestled in a bony structure called the sella turcica. The gland is attached to the hypothalamus by the pituitary stalk, around and through which course the veins of the hypophyseal-portal plexus. In most species the gland is divided into three lobes: anterior, intermediate, and posterior. In humans the intermediate lobe does not exist as a distinct anatomic structure but rather remains only as dispersed cells. Despite its proximity to the anterior pituitary, the posterior lobe of the pituitary is functionally distinct and is an integral part of a separate neural structure called the neurohypophysis (see below *The posterior pituitary [neurohypophyses]: Neurohypophyseal unit*).

Pituitary lobes

The cells comprising the anterior lobe are derived embryologically from an extension of the roof of the pharynx, known as Rathke's pouch. While the cells appear to be relatively homogeneous (of the same type) under a light microscope, there are in fact five different types, each of which, except in pathological states, secretes the same hormone or hormones throughout its existence. The thyrotroph synthesizes and secretes thyrotropin (thyroid-stimulating hormone, or TSH); the gonadotroph, both LH and FSH; the corticotroph, corticotropin (also called adrenocorticotropic hormone, or ACTH); the somatotroph, somatotropin (also called growth hormone, GH); and the lactotroph, prolactin (PRL).

These hormones are proteins that consist of large polypeptide chains. Furthermore, the gonadotropins and TSH are glycoproteins in which there is linkage to carbohydrates known as polysaccharides. Each of these three hormones is composed of two glycopeptide chains; one of which, the alpha chain, is identical in all three hormones, while the other, the beta chain, differs in structure for each hormone, lending specificity for individual hormone action. As is the case in all protein hormones, hormones of the anterior pituitary are synthesized initially in the cytoplasm of the cell as larger, inactive molecules called prohormones, which are split into the active hormone molecules at the time of secretion into the circulation.

Alpha and beta chains

Hormones. *Thyrotropin.* Thyrotropin is also called thyroid-stimulating hormone (TSH). Thyrotropin-producing cells (thyrotrophs) make up about 10 percent of the anterior pituitary and are located mainly in the centre of the gland. Thyrotropin becomes attached firmly to receptors on the surface of the thyroid cells, forming thyroid follicles in the thyroid gland. Following binding, a complex train of events occurs so that preformed thyroid hormones are secreted and steps are set in motion for the synthesis of additional thyroid hormones. Thyrotropin exerts other pervasive effects. It stimulates the growth of thyroid cells and leads to increased blood flow through the gland. It also enhances the breakdown of thyroglobulin, a large thyroid-hormone-containing glycoprotein that is stored within the follicles of the thyroid gland.

The levels of thyrotropin in circulating fluids become elevated during thyroid hormone deficiency because there is no negative feedback inhibition of pituitary thyrotropin release by thyroid hormone. Elevated thyrotropin levels are found in other pathological states, including the presence of a thyrotropin-producing pituitary tumour. Low serum thyrotropin levels occur following damage to cells in the hypothalamus that produce thyrotropin-releasing hormone (TRH), following damage to the pituitary stalk, or, finally, following damage to the thyrotrophs themselves. Tests of increased sensitivity have made the measurement of thyrotropin in blood valuable in detecting subtle changes of both thyroid hyperfunction and hypofunction.

Levels of thyrotropin

Gonadotropins. Gonadotrophs, which amount to about 7 percent of all pituitary cells, secrete two hormones, luteinizing hormone (LH) and follicle-stimulating hormone (FSH), but not in equal amount. The rate of secretion varies widely at different ages and at different times in the menstrual cycle of the female. Secretion of LH and FSH is low before puberty in both sexes. After puberty, about five times more LH than FSH is secreted. During menstrual cycles there is a dramatic rise in both hormones at the time of ovulation (see below *The ovary*), and secretion increases as much as 15-fold following menopause.

In men FSH stimulates the development of spermatozoa, in large part by acting on special cells in the testes called Sertoli cells. In women FSH stimulates the synthesis of estrogens as well as the maturation of cells lining the spherical, egg-containing structures known as the Graafian follicles. In menstruating women, there is a preovulatory surge in FSH levels in the blood. Inhibin, a hormone secreted by the Graafian follicles of the ovary and the Sertoli cells of the testis, inhibits the secretion of FSH from the pituitary gonadotroph.

In men androgens (male hormones) are secreted by specialized cells called Leydig cells, a process stimulated by LH. In women a preovulatory surge of LH is essential for rupture of the Graafian follicle so that the egg can be discharged on its journey to the uterus. The empty follicle becomes filled with other, progesterone-producing cells, transforming it into a corpus luteum.

Leydig cells

When a disease process leads to encroachment on the cells of the pituitary gland, usually the first evidence of cell failure is in the gonadotroph. Thus, disappearance of menstrual periods may be the first sign of a pituitary tumour in the female. In the male the most common symptom of gonadotropin deficiency is impotence. Isolated deficiencies of both LH and FSH do occur, but only rarely. In a male, LH deficiency alone leads to the appearance of what has been described as a "fertile eunuch"; there is sufficient FSH present to permit the maturation of spermatozoa, but because of the LH deficiency the man has, nonetheless, many of the characteristics of a castrate. Tumours also can produce an excess of LH or FSH, and pituitary tumours that secrete only the nonspecific, hormonally inactive alpha unit of glycoprotein hormones are not rare.

Corticotropin. Corticotropin, also called adrenocorticotropin hormone (ACTH), is a segment of a much larger prohormone glycoprotein molecule called pro-opiomelanocortin, which is synthesized by pituitary corticotrophs. This prohormone is split into a number of biologically active polypeptide fragments when the secretory granule is discharged from the cell. Among these hormones are corticotropin, whose major action is to stimulate growth and secretion of the cells of the adrenal cortex; alpha- and beta-melanotropin (melanocyte-stimulating hormone, MSH), which increases pigmentation of the skin; beta-lipotropin (LPH), which stimulates the release of fatty acids from adipose tissue; a small fragment of ACTH thought to improve memory; and beta-endorphin, a polypeptide that has excited a good deal of popular as well as scientific interest (see below *The adrenal cortex*).

Beta-endorphins (along with the enkephalins, which are neuromodulators) were discovered when investigators postulated that, since opiates such as morphine bind firmly to cell-surface receptors, there must exist natural substances that do likewise and have a narcotic action. The endorphins and enkephalins are known, therefore, as endogenous (self-generated) opiates or opioids. They have

Endorphins and enkephalins

powerful painkilling properties. Beta-endorphins instilled in the spinal fluid are capable of alleviating otherwise intractable pain in cancer patients. It has often been observed that severely traumatized individuals, those in battle, for example, appear to be free of pain. This phenomenon is due to the simultaneous release of beta-endorphin along with corticotropin in response to the stressful stimulus of the injury. There have also been reports of children with endorphin-producing pituitary tumours who are highly insensitive to pain. In addition, the release of endorphin or enkephalin may account for the euphoria ("high") experienced by long-distance runners. Finally, there is evidence, not fully accepted, that endogenous opioids stimulate appetite. This is seen in rats and obese persons who have a rare disease called Prader-Labhart-Willi syndrome. In these instances, the appetite is diminished after the administration of a narcotic antagonist, such as naloxone.

Hyperplasia or adenoma of corticotrophs gives rise to the constellation of symptoms called Cushing's syndrome. A deficiency of corticotropin also occurs both as part of the multiple deficiencies of panhypopituitarism and as an isolated defect. The diagnosis of corticotropin deficiency is important because afflicted persons who are also subjected to stress can succumb to severe shock. Once frequently administered in the treatment of disorders including allergic states, collagen disorders, and autoimmune diseases, corticotropin has been largely displaced by a number of synthetic variants of adrenal steroids.

Growth hormone. Somatotrophs are plentiful in the pituitary, constituting 40 percent of the gland. They are located predominantly in the lateral lobes and secrete between one and two milligrams of growth hormone (GH; also called somatotropin) per day. Growth hormone stimulates growth, not only of bone but of essentially all the tissues of the body. In biochemical terms, growth hormone simultaneously stimulates protein synthesis in tissues and enhances the breakdown of fat to provide the energy for the stimulated growth. Growth hormone is also an insulin antagonist and, in susceptible individuals, can lead to elevated sugar levels in the blood and diabetes mellitus.

Effects of GH

While GH may act on tissues directly, much of its effect is mediated by way of stimulating the liver and other tissues to manufacture and release secondary hormones, called somatomedins, which partly mimic the action of insulin. During childhood, somatomedin levels in the serum rise progressively with age, with an accelerated increase occurring at the time of the growth spurt of puberty, followed by a reduction to adult levels.

Growth hormone secretion is stimulated by growth hormone-releasing hormone (GHRH; also known as somatocrinin) and is inhibited by somatostatin. There are prominent daily fluctuations in growth hormone secretion in normal individuals, with the largest increase occurring shortly after the onset of sleep. Again, this increase is most pronounced at the time of puberty. Growth hormone levels in the serum are elevated in individuals with tumours that produce growth hormone, and its levels are unresponsive to stimulation in states of malnutrition.

There are many causes of short stature or dwarfism (see below *Growth and development*) other than deficient growth hormone secretion; for example, chromosomal abnormalities, malnutrition (including poorly controlled diabetes mellitus), thyroid deficiency, and disorders of bone formation are all examples of dwarfism with normal GH secretion. Nonetheless, growth hormone deficiency is a fairly common cause of short stature. Perhaps most frequent is GH deficiency resulting from damage to the hypothalamus and pituitary during fetal development or at birth because of trauma, lack of oxygen, or any of a number of other causes. When damage to the hypothalamus or pituitary is mild, growth hormone deficiency may be the only detectable manifestation of a disease state because the somatotrophs are the most sensitive of the pituitary cells to injury. When all of the cells of the pituitary are severely damaged or destroyed the patient is said to have panhypopituitarism (leading to diminished function of the gonads, the thyroid, and the adrenal glands).

Midgets usually suffer from one of two forms of hereditary (familial) isolated growth hormone deficiency. In some families the deficiency is the result of underproduction of GHRH, in which case growth hormone secretion may be stimulated by infusion of GHRH. In others, the problem lies in the somatotrophs themselves when they become incapable of manufacturing growth hormone. Growth hormone levels also tend to fall in some aged persons who otherwise appear to be normal.

In other forms of dwarfism, the hypothalamus and pituitary function adequately, and the abnormality lies rather in the lack of response of body tissues. A well-studied example is that of the Laron dwarf. These children suffer from a hereditary disorder characterized by the inability of growth hormone to bind to specific receptors in the body's tissues; circulating GH levels are elevated but somatomedin levels remain low because GH, unable to bind to receptors, cannot stimulate somatomedin secretion. Another example is the African Pygmy, in whom there is a resistance to the administration of GH. This is caused by an unresponsiveness to somatomedin, which suggests that there is a defect in the somatomedin receptors.

Growth hormone alone cannot generate growth without an adequate supply of food, so that in states of malnutrition dwarfism occurs in the face of a mild elevation in growth hormone concentrations in the blood.

Finally, an example of the effect of emotional and environmental factors on growth is found in the condition known as psychosocial dwarfism. Such children suffer emotional deprivation from uncaring or abusive parents. Growth hormone levels are low but return to normal along with an increased rate of growth when the children are removed to a more supportive environment, only to have the cycle repeated when the child is returned to the custody of the parents. These victims tend to be withdrawn and apathetic. They have disrupted sleep and bizarre eating and drinking habits. All of these symptoms are dramatically reversed when the child is removed to compassionate care in a hospital or foster home.

Psychosocial dwarfism

An adult GH-deficient dwarf has the body proportions of a young child. Radiographs (X-ray pictures) of growing ends of bone also show growth retardation in relation to the patient's chronological age. These changes are not apparent at birth but appear some time within the first two years of life. Puberty is often delayed, but untreated individuals may be fertile and give birth to normal children. When it appears in adults, GH deficiency produces only subtle changes, with minor decreases in strength and in the density of bones.

Growth hormone-deficient dwarfs respond dramatically to injections of human growth hormone. Supplies of GH were greatly limited in the past because the only source was GH extracted from human pituitary glands obtained at autopsies. With the availability of human GH manufactured by recombinant DNA technology using bacteria, the supply is potentially unlimited. Most treated patients achieve normal height, but in some, particularly those with the hereditary inability to synthesize growth hormone, antibodies to the injected growth hormone may block the therapeutic action. There is evidence that children with "constitutional short stature," that is, children from otherwise normal families in whom short stature is the rule in the absence of disease, may also respond to GH treatment.

Administration of human GH

Excess levels of growth hormone are most often caused by a benign tumour (adenoma) of somatotrophs of the pituitary gland. Rarely, a tumour of the lung or the pancreatic islets produces GHRH, which stimulates normal pituitary somatotrophs to excess secretion when released into the circulation. Even more rarely is there excessive, ectopic production of GH by tumour cells that do not ordinarily synthesize GH. If hypersecretion of growth hormone occurs during childhood, growth progresses at an inordinately rapid rate to extremes, 8 feet, 11 inches in the case of the "Alton Giant." Giantism is rare because such individuals usually have all of the infirmities described below for acromegaly.

The term acromegaly refers to the enlargement of the distal parts of the body; there is, in fact, progressive enlargement of the hands, feet, chin, and nose. Most other organs also become enlarged. The presence of a pituitary

tumour causes severe headaches, and the pressure of the tumour on the optic chiasm causes visual defects.

The acromegalic patient has overgrown supraorbital ridges, enlarged nasal sinuses that give a sonorous quality to the voice, an overgrown jaw, spaces between the teeth, and an enlarged tongue. The skin thickens, producing a permanently furrowed brow. The enlarged fingers are no longer tapered and become spatulated.

Because the metabolic actions of growth hormone are antagonistic to those of insulin, some acromegalic patients develop diabetes mellitus and are subject to all of its complications. Other problems include elevated blood pressure, heart disease, and progressive arthritis. Finally, because some of these tumours produce prolactin as well as growth hormone, males may have enlarged breasts, and both sexes may show abnormal lactation (milk secretion). Acromegaly can be treated with a considerable degree of success with surgery, with X-ray therapy, and with drugs such as bromocriptine or a synthetic, long-acting somatostatin.

Prolactin. On the evolutionary scale, prolactin is an ancient hormone serving multiple roles in mediating the care of progeny (it has been called the "parenting" hormone). Prolactin is a large protein molecule synthesized and secreted from cells, the lactotrophs, which compose 20 percent of the anterior pituitary gland and are located largely in the two lateral portions. Unlike other anterior pituitary cells whose activities are stimulated by hypothalamic-releasing hormones, the major modulating influence on lactotroph secretion is the inhibitory effect of the neurotransmitter dopamine, which, in the case of prolactin, functions as a hypothalamic neurohormone.

The "parenting" hormone

In the female, the major action of prolactin is to initiate and sustain lactation. In a breast-feeding mother, tactile stimulation of the nipples and breast by the suckling infant blocks the secretion of hypothalamic dopamine into the hypophyseal-portal circulation. This results in a sharp rise in serum prolactin levels followed by a prompt fall once feeding has stopped. Prolactin also inhibits secretion of GnRH from the hypothalamus and blocks the action of gonadotropins on the gonads. Thus, high prolactin levels reduce fertility in the female, protecting the lactating woman from a premature pregnancy. This protection is not absolute, however. Prolactin secretion is stimulated by estrogens and by TRH. This action of estrogens, much diminished in men, causes the level of prolactin to be relatively high in women. Finally, prolactin secretion is also stimulated by stress and exercise.

Actions of estrogens

Prolactin deficiency occurs along with the loss of other pituitary hormones in patients with panhypopituitarism from any cause. A striking example is that of Sheehan's syndrome, in which the anterior pituitary gland of the pregnant woman, for reasons poorly understood, is partly or totally destroyed during or shortly after the woman gives birth. Characteristically, in such a woman, breast milk is never produced.

Abnormally increased prolactin secretion may have many causes, including any of the many disease processes that damage the pituitary stalk (interrupting the flow of the prolactin inhibitor dopamine from passing through the hypophyseal-portal circulation to reach the lactotroph) and a number of drugs (particularly those used for the treatment of mental disease, high blood pressure, and the relief of pain). The most frequent cause of abnormally high prolactin levels is a tumour of the lactotrophs, termed a prolactinoma. In a large minority of hyperprolactinemic patients, however, no cause is discernible, and they are said to have "idiopathic hyperprolactinemia."

Prolactinomas were once thought to be quite rare. With the advent, in 1971, of a sensitive test for measuring serum prolactin, however, it became evident that hyperprolactinemia was common and that prolactinoma was the most frequently occurring pituitary tumour. It can be found usually in young adult females with abnormal lactation (galactorrhea) and disappearance of menstruation (amenorrhea), loss of sexual desire, and an inability to conceive. Prolactinomas are five times less common in men but are usually larger because the symptoms, particularly impotence, are gradual in onset.

Frequency of prolactinomas

In both sexes, symptoms attributable to the tumour mass alone, that is, headache and visual field defects, also occur. In women estrogen levels are decreased, resulting in osteoporosis. In men testosterone levels are lowered, contributing to a loss of physical strength as well as to impotence.

Initially, patients with prolactinomas were treated with X-ray therapy or neurosurgery; however, these forms of therapy largely have been replaced by the administration of potent drugs that mimic the neurotransmitter action of dopamine. These drugs promptly reduce elevated prolactin levels in all hyperprolactinemic patients, regardless of cause. In addition, individuals with prolactinoma usually demonstrate, sometimes quite strikingly, a decrease in the size of the tumour. This more conservative, pharmacological approach to treatment has been strengthened by the finding that patients with small prolactinomas may do well and exhibit no further tumour growth or increases in serum prolactin levels when left untreated.

THE POSTERIOR PITUITARY (NEUROHYPOPHYSIS)

Neurohypophyseal unit. The posterior pituitary lobe consists largely of extensions of processes (axons) from large clusters of cell bodies called nuclei (Figure 11). One pair, known as the superoptic nuclei, lies immediately above the optic tract, while the other pair, the paraventricular nuclei, lies on each side of the third ventricle of the brain. This anatomical complex forms the neurohypophyseal unit. There are neural connections upward to other centres of the brain, including a centre that modulates thirst. The two major neurohypophyseal hormones, vasopressin (also called antidiuretic hormone [ADH]) and oxytocin, synthesized in the cell body of the nuclei, descend through the long axons to be stored in secretory granules in the posterior lobe of the pituitary. Functionally, therefore, the posterior lobe is a storage and secretion site only.

Oxytocin and vasopressin. Oxytocin and vasopressin evolved from a single, primordial neurohypophyseal hormone, vasotocin, which is still present in lower vertebrates.

Evolutionary origins

After C.R. Kleeman in L.J. DeGroot *et al.* (eds.), *Endocrinology*, vol. 1, (1979); Grune and Stratton, Inc.

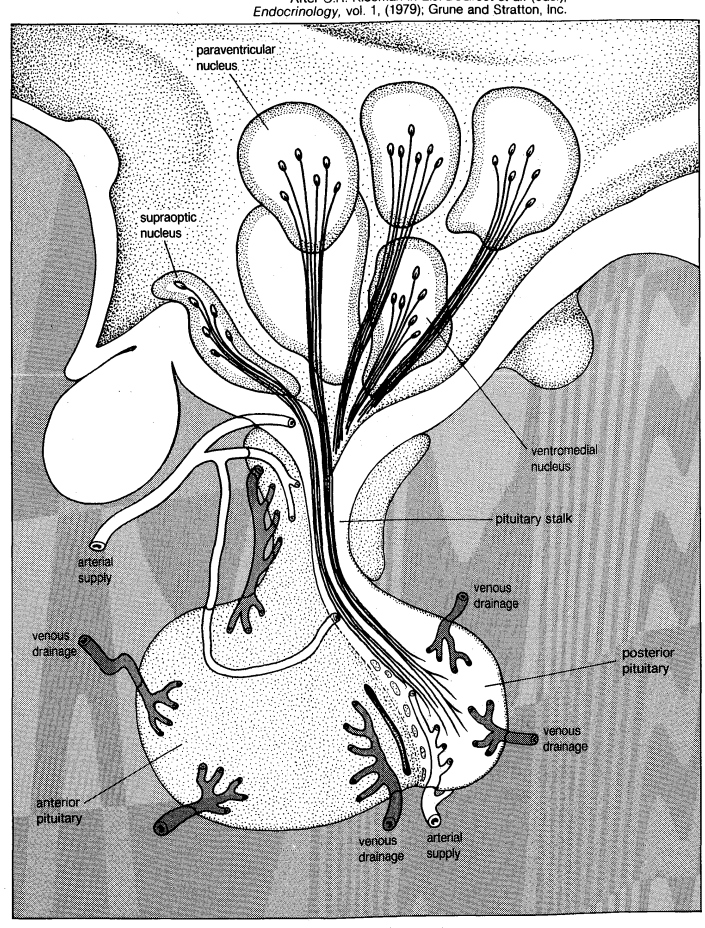


Figure 11: The neurohypophyseal unit.

Within the secretory granule, each hormone is attached to a large carrier protein, called neurophysin, from which it separates when the granule is discharged into the bloodstream. While vasopressin and oxytocin remain the major hormones synthesized by the neurohypophysis, it has become clear in recent years that other neuropeptides, including somatostatin, TRH, GnRH, CRF, and endogenous opiates are synthesized as well.

Vasopressin plays a key role in maintaining a constant total volume of water in the body and also in maintaining within narrow limits the concentration of dissolved substance (the osmolality) in those body fluids located outside body cells (extracellular water). In 1849, Claude Bernard noted that a small needle prick in the base of the brain led to a permanent, greatly increased urinary output. Following this, there ensued from the scientific community a confusing mixture of valid observations and false leads, until in the early 1930s Ernest Basil Verney showed unequivocally that the injection of a highly concentrated (hypertonic) salt solution into the carotid artery (the major artery carrying blood to the brain) resulted in a prompt increase in the excretion of urine from an animal's kidneys. This demonstrated (1) that there was a factor in the brain (vasopressin), which when let loose into the circulation enhanced water output, and (2) that this hormonal activity was stimulated by an increase in osmolality. It is now also known that vasopressin secretion increases in response to pain or stress.

Osmoreceptors

There is an osmoreceptor in the hypothalamus, which, when activated, leads to the release of vasopressin. Similarly, there exist two structures that are highly sensitive to distension and, in effect, serve as receptors for monitoring the total amount of fluid within the circulating blood: one is the carotid sinus, which is high in the neck and intimately associated with each carotid artery; and the other is a grouping of specialized cells in the left atrium of the heart. When the tissues of these structures are stretched by an expanding blood volume, nerves from these receptors carry impulses to the hypothalamus, thus inhibiting the cells of the neurohypophysis, shutting off the secretion of vasopressin, and resulting in increased urinary excretion of water.

The role of oxytocin is important, but more limited in scope. Oxytocin stimulates the contractions of the uterus, which are ongoing during the birth process; injections of oxytocin are used by obstetricians to stimulate uterine contractions in women whose labour is flagging. Oxytocin also prompts the milk glands of the mother's breast to release milk (milk let-down) within seconds after an infant begins to suckle by stimulating the contraction of muscular elements in the vicinity of the milk-containing glands.

There are no known diseases due to under- or overproduction of oxytocin. Emotional influences can affect oxytocin secretion; milk let-down may be premature, stimulated only by the cry of a hungry baby. While oxytocin is used to stimulate labour, delivery still may be normal in women in whom oxytocin deficiency is present.

Diabetes insipidus and inappropriate secretion of vasopressin. The clinical manifestations of diseases of the posterior pituitary may be considered in the context of two extremes in body water content: water intoxication (overhydration) and dehydration.

Water intoxication occurs when the body's ability to dispose of fluid is overcome by a large fluid intake or when the plasma volume percentage of water is increased because of defective mechanisms for the disposal of excess fluid, as is the case when more vasopressin is secreted than the body needs. Water intoxication from excessive fluid intake occurs rarely, having been reported in psychotic individuals, the winners of water-drinking contests, and individuals who have indulged in the overconsumption of beer (beer potomania).

A person becomes dehydrated when deprived of fluids or when there is excessive fluid loss from the body, such as occurs from excessive sweating, vomiting, or diarrhea. In these circumstances, the volume of fluid in the plasma is reduced and the concentration of solutes (the osmolality) is therefore proportionately increased. The decrease in body fluid and the proportional increase in solutes serve as potent stimuli for the secretion of vasopressin, which then acts on the kidneys to minimize urinary losses of water.

There are three disease states in which this regulatory mechanism fails. The first is termed adipsia (or hypodipsia), a rare disorder in which the brain's thirst centre is damaged. Individuals afflicted with adipsia become dehydrated, with little or no feeling of thirst. The problem can be alleviated by instructing them to drink adequate quantities of fluids at measured intervals.

Diabetes insipidus

The second disease, and by far the most common, is diabetes insipidus, so named because of the large volume of urine (which is tasteless rather than sweet as is the case in diabetes mellitus, where large quantities of the sugar glucose are present in the urine). Diabetes insipidus may be caused by trauma, including brain surgery, damage from brain tumours, or granulomatous infiltration, as in sarcoidosis, or, occasionally, for no discernible reason. Diabetes insipidus is rarely hereditary. Despite the frequency of head trauma, diabetes insipidus is an uncommon complication largely because it does not manifest itself until more than 85 percent of the neurohypophysis is destroyed.

The symptoms of diabetes insipidus include large urine volumes (usually from two to six litres each day, although up to 18 litres per day has been recorded) associated with frequent thirst and the ingestion of large quantities of water. If fluids are freely available the patient remains well except for the inconvenience of frequent drinking and of insomnia due to frequent urination. Occasionally, as a result of a patient's ongoing reluctance to urinate at frequent intervals, dilation of the kidney pelvis (hydronephrosis) and ureters (hydroureter) will occur, subsequently damaging kidney function. In the absence of a source of fluid, the patient becomes irritable and stuporous and will ultimately lapse into a coma and die. A highly satisfactory treatment is a long-acting, chemically modified form of vasopressin called desmopressin.

Nephrogenic diabetes insipidus

The third deficiency disease, a variant of diabetes insipidus, is called nephrogenic diabetes insipidus. It is a hereditary disorder linked to the X chromosome; males exhibit the disease whereas females are affected only slightly but are the sole carriers. The cause of the illness is not a deficiency of vasopressin (serum vasopressin levels may even be elevated); rather, the kidney tubules are defective and do not respond properly to the presence of vasopressin. Treatment with vasopressin or desmopressin is ineffective, but patients respond well to adequate fluid intake and a reduction in salt consumption.

The syndrome of inappropriate antidiuretic hormone secretion (SIADH) may be acute and life-threatening, characterized by sleepiness that progresses to convulsions, coma, and death, or, more commonly, chronic, in which the onset is far slower and is associated with few or even no symptoms.

Tumours of the neurohypophysis that secrete excess amounts of vasopressin have not been observed; however, other tumours, particularly those of the lung, may secrete large amounts of vasopressin, producing SIADH. For reasons not understood, any tumour that occurs in the brain may be associated with SIADH, and the syndrome has been noted in patients who have a wide variety of lung diseases. Finally, certain drugs, particularly chlorpropamide (used in the treatment of diabetes mellitus), that augment the action of normal amounts of secreted vasopressin may produce symptoms of SIADH.

The syndrome involves a lower than normal concentration of salt in the circulating fluid. Treatment of the acute form of SIADH involves the administration of concentrated salt solutions along with a powerful diuretic, so that the concentration of solutes is increased while the total plasma volume is decreased. The chronic form is satisfactorily treated with a drug called demeclocycline.

THE THYROID GLAND

All animal life requires oxygen for sustenance, and the human species is no exception. Oxygen drives the basic metabolic processes that permit growth, development, reproduction, physical movement, and constant body temperature. The complex of chemical interactions necessary to sustain these processes is called metabolism, and the

prime, overall regulators of metabolism are the thyroid hormones.

Anatomy. The thyroid gland is located in the anterior part of the neck in the midline. It consists of two lateral lobes lying on each side of the thyroid cartilage (Adam's apple) and connected by a band of tissue called the isthmus. It is one of the larger endocrine glands, and its capacity to grow is phenomenal. Any enlargement of the thyroid, regardless of cause, is called a goitre. The thyroid arises in the embryo from a downward outpouching of the floor of the fetal pharynx, and a persisting remnant of this migration is known as a thyroglossal duct.

Goitre

If viewed under a three-dimensional microscope, the resting thyroid is seen as a collection of small, generally globular sacs, called follicles, filled with the prohormone thyroglobulin. The cells lining these globules are called follicular cells, and it is their function to synthesize thyroid hormones as part of the prohormone thyroglobulin and either to secrete them directly into the circulation or store them within the follicles. When the individual's requirement for thyroid hormone increases, thyroglobulin is split into its component parts, and the thyroid hormone thus released passes through the follicular cells to enter the circulation. Nestled in the spaces between the follicles are parafollicular cells. These, in essence, form a separate endocrine organ. They have an entirely distinct embryological origin, and they are not embedded in the substance of the thyroid gland, in many species other than man (see below *The parathyroid glands: Calcitonin*).

Biochemistry. The thyroid hormones are not proteins; rather, they are modifications, called thyronines, of an amino acid, tyrosine. Thyroid hormones are heavily laden with iodine. The major active thyroid hormones are thyroxine (T_4) and triiodothyronine (T_3). Even though the thyroid gland manufactures considerably more T_4 than T_3, T_3 is roughly $2\frac{1}{2}$ times more potent than T_4. Indeed, in many ways, T_4 serves as an additional, circulating depot for T_3 in that when T_4 leaves the circulation and travels through the cytoplasm to the nucleus of the target cell, its action at that site is preceded or accompanied by its conversion to T_3.

Thyroid hormones

Most of the T_4 and T_3 secreted by the thyroid is bound to special proteins (thyroxine-binding globulin [TBG] and prealbumins) in the serum, although small amounts of these hormones travel freely in the serum and are readily taken up by tissues to be replenished instantaneously from the T_4 that had been attached to the binding proteins.

Essentially all the cells in the body are target cells of thyroid hormones. The major function of the thyroid hormones is to stimulate the synthesis of protein once they have entered the cell nucleus. Another important function is to stimulate the activity of the cell's mitochondria. These intracellular organelles are the sites at which there is a controlled exchange of energy. Some energy is conserved for the body's functionings, while the remainder is dissipated as heat. The proportion of energy devoted to each of these processes is controlled by the thyroid hormones. There are other intracellular thyroid hormone functions that are not well understood, but it is clear that thyroid hormones modulate protein, carbohydrate, fat, and vitamin metabolism, as well as the generation of body heat. Thyroid hormones also modify the activity of the autonomic nervous system.

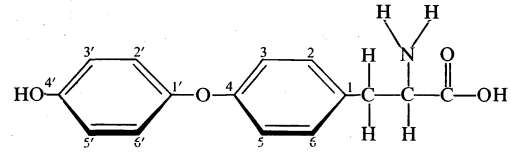
Regulation of hormone secretion. While multiple factors, including nerves supplying the thyroid gland, influence thyroid hormone secretion, by far the major influences are the negative feedback loops. The thyroid is a prime example of the negative feedback effects of the hypothalamic-pituitary-target organ axis. Briefly, thyroid hormones inhibit the release of thyrotropin-releasing hormone (TRH) from the hypothalamus and thyrotropin (thyroid-stimulating hormone [TSH]) from the anterior pituitary. Increased consumption of thyroid hormones decreases their concentration in the circulating fluids, resulting in enhanced thyrotropin secretion and thus an increased thyroid hormone secretion until a normal serum level is regained. Conversely, with the administration of the thyroid hormones, the resultant increased serum levels inhibit TRH and thyrotropin secretion and reduce the secretion of thyroid hormone from the thyroid gland until the elevated circulating thyroid level is returned to normal. If an amount of thyroid hormone equal to the normal daily thyroid output is administered to a patient, the thyroid gland is effectively suppressed; it produces no thyroid hormone because levels of circulating TSH are greatly reduced.

Major regulatory influences

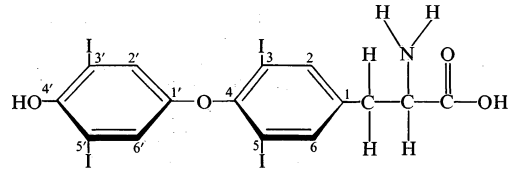
There is an important extrathyroidal mechanism for modulating thyroid hormonal activity, that is, the controlled conversion of T_4 into either the potent hormone T_3 or the inactive molecule rT_3 (Figure 12). Tissue enzymes, particularly abundant in the liver and kidney, control the conversion of T_4 to T_3 or reverse triiodothyronine (rT_3). Consequently, when T_4 is metabolized to T_3, thyroid hormone action is enhanced. Similarly, when the pathway for the conversion of T_4 to rT_3 is favoured, T_3 levels fall and thyroid hormone activity in that particular tissue is proportionally decreased.

Aside from the regulatory functions, other factors, external or internal, may also influence the circulating levels and utilization of thyroid hormones. In all forms of malnutrition, including anorexia nervosa, there is a significant reduction in the conversion of T_4 into T_3. The commensurate decrease in oxygen consumption and metabolic rate has survival value for a person deprived of adequate food to sustain health; in effect, death from starvation is postponed. Iodine intake is important because an inadequate dietary supply leads to reduced circulating thyroid levels and an ensuing increase in serum thyrotropin levels. This increase, while perhaps not adequate to produce sufficient thyroid hormone, nevertheless stimulates growth of the thyroid, with the resultant appearance of a goitre. In the short term, low environmental temperature leads to in-

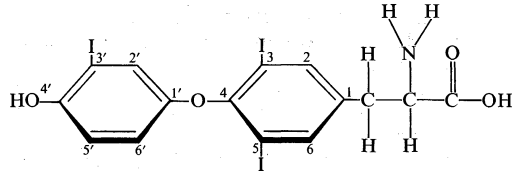
External and internal influences



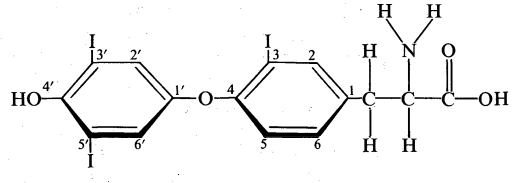
thyronine



3,5,3′,5′-tetraiodothyronine (T_4)
(thyroxine)



3,5,3′-triiodothyronine (T_3)



3,3′,5′-triiodothyronine
(reverse T_3; rT_3)

Figure 12: Structural diagrams of T_3, rT_3, and T_4, showing the synthesis of T_3 and rT_3 from T_4.

creased utilization of thyroid hormones, activation of the hypothalamic-pituitary-thyroid axis, and a consequent rise in T_4 and T_3 levels. As the environmental temperature rises, the converse results, and small, appropriate changes in normal persons have been noted with changes of season.

Finally, thyroid hormone levels may be affected by many illnesses that have nothing directly to do with the thyroid gland. For this reason, it is not easy to ascertain with certainty the influence of aging on thyroid hormone activity because it is difficult to accumulate large numbers of aged subjects who can be said to be free of any disease. It is generally agreed, however, that few important changes occur in thyroid activity during the normal aging process.

Diseases and disorders. *Hyperthyroidism.* When the human body is exposed to excessive amounts of thyroid hormone the result is a disease known as hyperthyroidism (or thyrotoxicosis). The most common cause of hyperthyroidism is Graves' disease, named after the Irish physician Robert J. Graves, who was among the first to describe it. It is noteworthy that hyperthyroidism is at least seven times more common in women than in men, although the reasons for this are poorly understood. Because there is a complex, hereditary tendency for the production of thyroid autoantibodies, it is not rare for Graves' disease to occur in many family members.

Graves' disease

Hyperthyroidism typically begins with a gradual onset of a constellation of symptoms, including increased nervousness and emotional instability associated with a fine tremor of the hands. The patient feels warm, perspires freely, and is intolerant of heat since a greater proportion of the body's energy is dissipated as heat. The pulse races, the heart thumps, and the systolic element of the blood pressure is elevated. In severe cases of hyperthyroidism, heart failure may occur. The drive to physical overactivity is dampened by increasing weakness and easy fatigue. Bowel movements may be normal, but they are often punctuated by bouts of diarrhea. Menstrual periods may become scant or may disappear entirely. Perhaps most striking is the apparent paradox of an increase in appetite associated with a loss of weight, the result of the fact that the excess of thyroid hormones leads to an increased metabolism. There is often, but not always, a swelling in the neck, and the physician's fingers may often detect the outline of an enlarged thyroid gland.

In patients with Graves' disease, an additional group of symptoms may appear. The eyes protrude (exophthalmos) and the distance between the opened eyelids increases so that in extreme cases the eyes may not close completely. The eyeball muscles and the entire orbit becomes inflamed and the patient may complain of double vision. Less often, there is a thickening of the skin over the shins (localized myxedema) and of the skin of the fingers (clubbing).

These changes, along with hyperthyroidism itself, stem from a pathological process called autoimmunity in which the body's immune system generates autoantibodies, called immunoglobulins, that are harmful to the body's own tissues. The first and most common cause of hyperthyroidism is the presence of antibodies producing Graves' disease. At least three sets of antibodies are involved. In the genesis of thyroid hypofunction, there is a fourth set of antibodies that competes with the hormones but does not induce the actions of the stimulating hormones once they are bound to the receptors.

The first set of antibodies is called the thyroid-stimulating immunoglobulins (TSI), which exert extraordinary effects unique throughout the endocrine system. They have a molecular configuration that mimics that of thyrotropin so that they are attracted to, and bind tightly with, the same receptors on the surface of thyroid cells that attract thyrotropin. The effects of thyrotropin are mimicked exactly. The result of TSI stimulation is the same as that of thyrotropin stimulation: thus, the number of follicular cells multiply and the thyroid enlarges, the follicles empty as the prohormone thyroglobulin is split, thyroid hormones are released, and the follicular cells are stimulated to synthesize and secrete excessive amounts of thyroid hormone. The end result is hyperthyroidism, or thyrotoxicosis.

TSI stimulation

The second set of autoantibodies attack orbital contents, including the eyeball muscles, producing Graves' ophthalmopathy and, less frequently, localized myxedema (dry, waxy swelling of the skin). The third set of antibodies is cytotoxic; that is, they damage and eventually destroy follicular cells. These cytotoxic immunoglobulins are an important cause of hypothyroidism, but when present early in the course of Graves' disease, they may only limit the effect of TSI. Thus a patient may manifest ophthalmopathy with normal or even reduced thyroid function. Furthermore, cytotoxic antibodies, even in the absence of TSI, may acutely damage thyroid follicles, leading to a leakage of large quantities of thyroid hormones so that a thyrotoxic state ensues.

The second most common cause of hyperthyroidism is toxic multinodular goitre. It begins early in life with iodine deficiency or other factors that block thyroid hormone secretion; the resulting low T_4 and T_3 levels lead to unremitting TSH secretion and to constant thyroid gland stimulation. This, in turn, produces glandular enlargements and the eventual formation of multiple nodules that produce excessive amounts of thyroid hormones autonomously. Less common is a benign tumour (toxic adenoma) of the thyroid, and rarely a malignant tumour may hypersecrete thyroid hormones. In such patients, hyperthyroidism may be due to overproduction of hormones from a metastatic deposit located in one or more parts of the skeleton, even though the thyroid gland itself has been removed surgically. Another rare form of hyperthyroidism results from inappropriate thyrotropin secretion. This may be caused by increased thyrotropin secretion resulting from a tumour of the pituitary thyrotrophs. Another cause of thyrotropin overproduction and secretion is the loss of cell receptors for thyroid hormones on the surface of thyrotrophs. As a consequence of this loss, the pituitary is not inhibited from releasing thyrotropin, resulting in persisting thyrotropin release.

Hyperthyroidism may also result from a hyperfunctioning ectopic tumour of thyroid tissue in the ovary (struma ovarii) or from the ingestion of excessive amounts of thyroid hormones. On occasion a person will, in order to lose weight or for some other reason, take large, toxic unprescribed amounts of thyroid hormone and may persist even to the point where surgical removal of the thyroid gland becomes necessary.

Effective treatments for hyperthyroidism include (1) surgical removal of all but a remnant of the thyroid gland, (2) the administration of drugs that specifically block the synthesis and release of thyroid hormones, and (3) the administration of radioactive iodine. This last form of treatment is effective because the thyroid gland, unable to distinguish between stable and radioactive forms of iodine, extracts both from the serum. Thus, follicular cells are damaged by the concentration of radioactivity within them to the point that the hyperthyroid state is relieved and the thyroid function returns to normal. In some instances, as is the case in excessive hormone release from inflammation of the thyroid or following ingestion of large amounts of thyroid hormone, drugs that block the manifestations of thyroid action on tissues, such as propranolol, are effective. Finally, for reasons not fully understood, using stable (nonradioactive) iodine also impairs release of thyroxine from the gland: improvement, however, may be short lived.

Treatments for hyperthyroidism

Hypothyroidism. Like hyperthyroidism, hypothyroidism can have many causes. It is, however, less common. In fact, overtreatment of hyperthyroidism with either radioiodine or surgery has emerged as the most frequent cause of hypothyroidism. In other instances, however, a child is born without a thyroid or an adult becomes hypothyroid without apparent cause. In some cases, autoantibodies appear and bind to thyrotropin receptors on the follicular cell, but unlike TSI, these autoantibodies are not agonists and do not accelerate the secretion of thyroid hormones. Instead, they are antagonists because they block access of thyrotropin to the receptor. As a result of their actions, the thyroid gland atrophies. Hypothyroidism may also occur as a result of disease of the cells of the hypothalamus that produce TRH or of the thyrotrophs of the anterior pituitary.

Hypothyroidism may also be associated with an enlarged

Hashimoto's thyroiditis

thyroid, a goitre. This is most commonly caused by inflammation of the thyroid (Hashimoto's thyroiditis) due to cytotoxic autoantibodies (see above *Thyroid gland: Hyperthyroidism*). The thyroid enlarges because of a heavy infiltration of white blood cells called lymphocytes. As discussed above, iodine deficiency results in goitre formation because of constant stimulation from elevated TSH levels in the serum. When iodine deficiency is severe, these compensatory efforts are inadequate and the patient becomes hypothyroid. In rare families there appears a hereditary absence of one of the enzymes essential for the synthesis of thyroid hormones. Although such individuals are hypothyroid from birth, they develop large goitres because of constant stimulation of the thyroid by TSH, a condition referred to as goitrous cretinism.

A number of drugs can block thyroid hormone synthesis and thus lead to goitrous hypothyroidism. Among them are the antithyroid drugs used in the treatment of hyperthyroidism, and lithium, prescribed for psychiatric disorders. There are a number of naturally occurring vegetable goitrogens, particularly cabbage. Finally hypothyroidism may be due to the absence of tissue receptors for the thyroid hormones. Persons with this very rare disorder have high but ineffectual serum levels of T_4 and T_3.

The onset of hypothyroidism may be gradual and subtle, so that it is often missed not only by the patient but also by a physician. It may be mild and difficult to diagnose, or all of its symptoms and conditions, called myxedema, may be present. The term myxedema stems from the fact that, for reasons not well understood, the hypothyroid patient produces an excess of a thick protein-containing (myxomatous) fluid that is deposited in the skin and other organs.

In many instances, the hypothyroid patient shows symptoms diametrically opposed to those of the patient with thyrotoxicosis. The patient is sluggish in movement and thought and has a thick, dry skin with coarse, dry, thinning hair. The patient does not eat excessively but gains weight. (Excessively obese persons, however, are rarely hypothyroid.) The tongue is large and impedes articulation of the guttural voice. The eyes are puffy and the lids lowered. Reflexes are slow, and there is continuing weakness. Females often have excessive menstrual bleeding and are relatively infertile. Patients prefer hot weather and are intolerant of cold. In fact, those with myxedema cannot generate additional body heat in response to a cold environment, so that when exposed to extreme cold, their body temperature may fall to levels as low as 74° F (23.3° C).

Myxedema is relatively common in the elderly, and the symptoms and signs are often mistaken for changes attributable to old age. While it is true that every endocrinologic disorder, whether hyperfunction or hypofunction, has been found to be sometimes associated with a mental aberration, most often depression, this may be striking in severe hypothyroidism; it has been called "myxedema madness." Treatment with T_4 may return the patient to a normal mental state, but the mental illness may remain unchanged or in some instances become even worse.

The same myxomatous fluid that infiltrates the skin often accumulates in body cavities as well. Thus what appears to be an enlarged heart in a chest X-ray film may be a benign collection of fluid in the pericardium, and similar changes may be found in the pleural and abdominal cavities.

Cretinism

Since the thyroid hormones pass only poorly from the maternal to the fetal circulation, iodine-deficient fetuses or those without thyroid glands become hypothyroid in utero and are born as cretins (infants whose growth and mental development are arrested) with a characteristic appearance somewhat similar to that present in myxedematous adults. Hypothyroidism also may be produced in the fetus when a pregnant woman is exposed to radioactive iodine or antithyroid drugs. Cretinism is associated with severe mental retardation, so that it is essential that hypothyroid infants be treated promptly with thyroid hormone. Indeed, it has become routine to check thyrotropin and T_4 levels in all infants at birth so that a child with any degree of hypothyroidism can be identified and promptly given the appropriate treatment.

Treatment of the adult with myxedema is relatively simple. The patient is given enough thyroxine to increase serum T_4 and decrease serum thyrotropin to normal levels. While the mental retardation of infantile cretinism cannot be reversed, both children and adults can be returned to a state of normal physical health.

Thyroid tumours. Thyroid tumours are remarkable in two respects. First, patients exposed to radiation from any source (nuclear blast, radioactive iodine, or X rays) have a much increased risk of developing thyroid tumours, including thyroid cancers. Second, unlike many other organ cancers, the most common thyroid tumour, a papillary carcinoma, pursues a slowly developing, painless course compatible with a long life span, and it is held in check when enough thyroxin is administered to suppress thyrotropin secretion. Other, less common thyroid tumours pursue a much more threatening course, and the most malignant of these may cause death within a few months to a few years.

Diagnostic techniques have improved to the point that it is relatively easy to ascertain the nature of a lump (mass) found on the thyroid gland. An image of the accumulation of administered radioactive iodine or technetium in the thyroid (thyroid scan) will demonstrate whether the mass is "hot" or "cold," that is, whether the mass is functionally active or not, functional activity being rare in thyroid cancers. Similarly, imaging with ultrasound will reveal whether the mass is fluid (cystic) or solid, cancers being solid. Finally, cells obtained by suction through a fine needle inserted through the skin into the mass may be examined under the microscope. These methods, particularly the last, may obviate the need for exploratory thyroid surgery.

The treatment of thyroid masses, also known as thyroid nodules, has long been controversial, but with increasing awareness of the slow course that many of these tumours pursue, surgeons now employ less radical procedures in dealing with them.

THE PARATHYROID GLANDS

Importance of calcium

The level of calcium in the blood is closely regulated, and wide fluctuations in either direction can be life-threatening. Calcium is a key element in the human body. Not only does it serve as the major constituent for bone, but it is also essential for the normal functioning of all body cells, as it is a mediator for many cell functions. For example, without calcium, blood will not clot. Many of these actions also require adequate supplies of magnesium and phosphorus. A healthy body needs a regular, continuous supply of these elements: about a gram each day for calcium and phosphorus and about one-third as much for magnesium.

Almost all the calcium contained in the body is deposited in bone (about 1.3 kilograms in the normal adult). While this mass provides skeletal support and serves as a reserve from which calcium may be mobilized, it is the remaining 1 percent, dissolved in body fluids, whose concentration is so carefully monitored. In the plasma, calcium exists largely as a dissociated ion (Ca^{2+}) loosely bound to plasma proteins with a small proportion bound more tightly to phosphate and citrate. To insure that calcium levels and distribution are maintained within narrow limits, parathyroid hormone (PTH), calcitonin, and the calciferols (the active metabolites of vitamin D) serve regulatory functions.

Anatomy. The parathyroid glands, usually four in number, are small structures adhering to or even imbedded in the substance of the thyroid gland. It is not surprising, therefore, that they were recognized as distinct endocrine organs rather late in the history of endocrinology, first described by a Swedish anatomist, Ivar Sandström, in 1880. At the beginning of the 20th century, symptoms due to parathyroid deficiency were attributed to the absence of the thyroid since the surgical removal of one was frequently accompanied by the inadvertent removal of the others. In 1909 an American pathologist, William G. MacCallum, recognized that parathyroid deficiency could be mitigated by the injection of calcium salts, and not until 1925 was an active parathyroid extract prepared by a Canadian biochemist, James B. Collip. In 1925 an Austrian surgeon, Felix Mandl, was the first to remove a

parathyroid tumour from a patient, and thereafter this and related subjects were extensively explored by the American clinical endocrinologist Fuller Albright.

Embryological development

The parathyroids arise in the embryo from the third and fourth pairs of branchial pouches, bilateral grooves resembling gill slits in the neck of the embryo and reminders of man's evolutionary debt to fishes.

Hormones. *Parathyroid hormone.* The parathyroids produce only one major hormone, parathyroid hormone (PTH), also called parathormone. Under the microscope the PTH-producing cells, the chief cells, occur in sheets interspersed with areas of fatty tissue. Occasionally the cells are arranged in follicles, similar to but smaller than those present in the thyroid gland. In common with other endocrine glands, the parathyroids synthesize a large prohormone, which is inactive. At the time of secretion the prohormone is split into an inactive fragment and PTH (a polypeptide containing 84 amino acids).

In contrast to the elaborate mechanisms controlling the secretion of other endocrine glands, the major determinant of PTH secretion is the level of ionized calcium in the serum (see above *The nature of endocrine regulation*). Should the serum calcium level rise, PTH secretion is inhibited. Conversely, should it fall, PTH levels rise. Magnesium controls PTH secretion in a similar fashion.

Actions of parathormone

The actions of PTH are multiple but they are all geared toward raising the level of ionized calcium in the plasma. Parathormone mobilizes calcium from bone by stimulating the activity of large, bone-dissolving cells called osteoclasts. It acts on the kidney to enhance the reabsorption of calcium by kidney tubules so that excretion of calcium in the urine is reduced. Parathyroid hormone acting in concert with vitamin D metabolites also enhances the absorption of ingested calcium from the bowel, and there is evidence that it provokes the transfer of some calcium from the milk in the breast of a lactating woman into her blood. On the other hand, PTH is a powerful inhibitor of renal tubular reabsorption of phosphate. Finally, an ancillary action of PTH is to assist in the regulation of body acidity by blocking tubular reabsorption of bicarbonate.

Calcitonin. Calcitonin was not recognized as a specific hormone until 1962. Calcitonin is a polypeptide containing 32 amino acids. It is synthesized and secreted from cells, termed parafollicular, or C, cells, which lie between the follicles of the thyroid gland. These cells do not have the same embryological origin as do the thyroid follicular cells; they migrate into the substance of the thyroid from a fetal structure called a branchial pouch. Human calcitonin differs considerably from the calcitonin of other species, and physicians take advantage of these differences when they administer salmon calcitonin, which provides a longer lasting, more potent action than does human calcitonin.

Major action

The major action of calcitonin is to lower the level of calcium in the blood by sharply inhibiting the ongoing dissolution of calcium from bone. Not unexpectedly, calcitonin secretion is stimulated whenever serum calcium levels rise above the normal range so that, between them, calcitonin and PTH effectively maintain steady calcemia in a normal individual.

Vitamin D and the calciferols. Unlike calcitonin, the awareness of vitamin D is relatively ancient. Vitamin D deficiency was first described more than 300 years ago as rickets, but it was not until 1971 that the chemical transformations that make vitamin D biologically active were described. The term vitamin D refers to a family of compounds that are derived from cholesterol. There are two major forms of vitamin D: vitamin D_3, found in animal tissues and often referred to as cholecalciferol, and vitamin D_2, found in plants and now better known as ergocalciferol. Both of these compounds are inactive precursors of potent metabolites; they fall, therefore, into the category of prohormones. This is true not only for the cholecalciferol found in animal tissues but also for that which is generated in human skin following exposure to ultraviolet light. These precursors are modified during their passage through the liver to a sterol called 25-hydroxycholecalciferol, and then further modifications, modulated by the serum PTH level, occur in the kidney. One of these products, 1,25-dihydroxycholecalciferol (calcitriol), is the most potent derivative of vitamin D. The other, 24,25-dihydroxycholecalciferol, has actions that are not clearly defined at present.

Vitamin D deficiency

Persons with a vitamin D deficiency suffer from rickets, characterized by soft, poorly calcified bone, along with poor absorption of calcium. Calcitriol or any of its precursors promotes a dramatic increase in the absorption of calcium by the intestine and a prompt repair of the diseased bone. It is generally agreed that the improvement in the bone results from the alleviation of the calcium deficiency; calcium is resorbed, but bone synthesis is not enhanced.

Diseases and disorders. *Hyperparathyroidism.* Overactivity of the parathyroid glands was originally thought to be a rare disorder because it was generally considered only in those patients who had definite symptoms. This view changed precipitously when the measurement of multiple plasma constituents, including calcium, became an integral part of a routine health examination. Hypercalcemia (excessive levels of calcium in the bloodstream) associated with few or no symptoms occurs in one in 1,000 adults, representing a large subpopulation in Western countries.

While there are many other disorders associated with elevated levels of calcium in the serum, including malignancy and the ingestion of too much vitamin D, primary hyperparathyroidism (primary in the sense that the parathyroid hyperfunction is not due to a known cause) is the preeminent cause of hypercalcemia. In hyperparathyroid patients, the hypercalcemia is often accompanied by a reduction in serum phosphorus levels and an increase in the levels of serum uric acid and serum acidity.

Almost all the symptoms of hyperparathyroidism result from hypercalcemia, but not all hypercalcemic patients become ill. (Thus, it is important for the physician to distinguish hyperparathyroidism from a chemical anomaly, called familial hypocalciuric hypercalcemia, in which elevated serum calcium levels are associated with a reduction in urinary calcium excretion. This condition is benign and usually no treatment is required.) Primary hyperparathyroidism results most often from an adenoma (a benign tumour), which secretes an excessive amount of PTH despite the elevation in serum calcium that it produces; because the adenoma is autonomous and not subject to negative feedback loops, elevated serum calcium levels are not followed by inhibition of PTH secretion. Primary hyperparathyroidism occasionally is associated with parathyroid hyperplasia, an increased number of hyperfunctioning cells that do not, however, cluster to form a typical adenoma. Rarely, a malignant tumour (a carcinoma of the parathyroid gland) may produce extraordinarily large amounts of PTH, which, in turn, produce dangerously high levels of serum calcium.

Asymptomatic hypercalcemia

With the advent of screening tests, large numbers of persons with mildly elevated serum calcium levels have been identified although the majority of these individuals are without symptoms. This form of asymptomatic hypercalcemia occurs most frequently in postmenopausal women. Symptoms include weakness, loss of appetite and weight loss, nausea, vomiting, and mental depression. There may be increased urinary output with an increased thirst and fluid intake. Constipation is a frequent problem. With severe, rapidly progressing hyperparathyroidism, there may be bone pain, stupor, and even coma. Weakened bones may form cysts (osteitis fibrosa cystica) and may break after little or no physical stress (pathological fractures). Since calcium does not dissolve readily in serum, elevated levels result in a precipitation of calcium deposits in susceptible tissues, most prominently the kidney, and the pain of kidney stones (renal colic) is often the first evidence of hyperparathyroidism. Kidney damage may progress to the point where the patient's life is threatened.

Although primary hyperparathyroidism may be hereditary or in some instances associated with multiple endocrine neoplasia (see below *Ectopic hormone and polyglandular disorders*), most often the cause of primary hyperparathyroidism is unknown. Known causes, referred to as secondary hyperparathyroidism, usually involve an unrelated kidney disease. When kidney failure occurs, serum calcium levels fall. The resulting increase in PTH

secretion often leads to severe bone disease along with intractable itching.

Causes of hypercalcemia

Hypercalcemia may result when malignant tumours (particularly of the lung) secrete substances, in most instances not PTH, that increase the rate of dissolution of bone. Another important cause of hypercalcemia is vitamin D intoxication (discussed below). The most common cause of hypercalcemia other than primary hyperparathyroidism results from invasion and destruction of bone by the spread of a cancer, most commonly cancer of the female breast. A number of drugs, most prominently diuretics such as hydrochlorothiazide or furosemide, may also increase serum calcium levels.

The treatment of symptomatic hyperparathyroidism is surgical removal of the tumour. The treatment of asymptomatic hyperparathyroidism is less clear-cut. Because some patients may remain symptom-free for years, one alternative is simply to observe the patient's course without treatment unless symptoms appear. If continued observation alone is psychologically distressing, however, surgical removal of the tumour is warranted. In the case of mild postmenopausal hyperparathyroidism, treatment with the estrogen hormone estradiol is effective for many patients.

If the serum calcium rises to dangerous levels, it can be lowered quickly by using intravenous fluids with a powerful diuretic, thus "washing out" the excess calcium. The drug plicamycin (mithramycin) is highly effective in lowering serum calcium, although it may have toxic side effects.

Hypoparathyroidism. If PTH secretion is greatly reduced or ceases entirely, mobilization of calcium from bone and other sources ceases, and a fall in serum calcium to abnormally low levels results. Hypoparathyroidism is a rare disorder; indeed, the most common cause is iatrogenic—*i.e.*, physician- or treatment-induced, such as the PTH deficiency that occurs following the inadvertent removal of parathyroid glands during thyroid surgery. Spontaneously occurring hypoparathyroidism is probably an autoimmune disease because serum autoantibodies are found in some afflicted individuals. This form of hypoparathyroidism may appear in the syndrome of multiple endocrine deficiencies (see below *Ectopic hormone and polyglandular disorders*). Impaired PTH secretion may occasionally occur in the presence of intact parathyroid glands. Such is the case when a person suffers from magnesium deficiency, usually associated with alcoholism. In such patients serum calcium levels remain persistently low until the magnesium deficiency is repaired. Finally, Albright described what he termed pseudohypoparathyroidism in which there is a defect in the binding of PTH to its cell surface receptor.

Symptoms of hypoparathyroidism

The symptoms of hypoparathyroidism are essentially those resulting from low levels of serum calcium. Most prominent is muscular cramping and twitching, exemplified dramatically by carpopedal (wrist and foot) spasms; during the spasms there are painful cramps of the toes and feet, along with severe, tetanic contractions of the muscles of the hands so that the four fingers are rigidly extended while the thumb presses against the palm. This neuromuscular excitability can progress to generalized convulsions. In addition, patients with long-standing hypocalcemia develop cataracts and calcification in the basal ganglia of the brain, which in turn can produce symptoms of parkinsonism. Occasionally, patients also have a spotty depigmentation of the skin (vitiligo) and hair loss. Patients suffering from pseudohypoparathyroidism also may have peculiar skeletal abnormalities: "short coupled," with a short neck and extremities, obese, with a rounded face, and sometimes shortened metacarpal bones.

When treatment is urgent, the patient is given calcium salts intravenously. Long-term therapy consists of treatment with vitamin D or one of its metabolites, along with calcium salts by mouth. Serum calcium levels must be monitored to be certain that, on the one hand, the patient is given enough medication to avoid hypocalcemia and, on the other hand, to prevent the hazards of hypercalcemia with its attendant complications, such as kidney stones.

It should be noted that there are causes for hypocalcemia other than hypoparathyroidism. In the past, vitamin D deficiency (rickets) was a common cause, but with the wide distribution of vitamin D supplements in milk and other foods this has become a rare event. There remain, however, patients who suffer from abnormalities in the metabolism of vitamin D (vitamin D-resistant rickets), which may be treated effectively either with very large doses of vitamin D or with 1,25-dihydroxycholecalciferol. Severe inflammation of the pancreas (pancreatitis) is associated with hypocalcemia, and low serum calcium levels may also occur in patients who suffer from intestinal malabsorption (sprue). In these individuals, ingested calcium binds to unabsorbed fat and is excreted. Treatment of the underlying condition relieves the symptoms.

Hypercalcitoninemia. It was not until 1968 that tumours of the parafollicular cells of the thyroid gland were discovered to secrete large amounts of calcitonin. These tumours sometimes occur among family members and sometimes as isolated cases. Such tumours, known as medullary carcinomas of the thyroid, also occur in one of the forms of multiple endocrine neoplasia (see below *Ectopic hormone and polyglandular disorders*). In most patients, serum calcium levels are not low, as might be expected, because any tendency to hypocalcemia is countered by increased PTH secretion. The threat of medullary carcinoma is the fact that it invades local areas in the neck and spreads to distant organs, resulting in death. Patients with these tumours have elevated serum calcitonin levels or are hyperresponsive to stimulation of the parafollicular cells by an infusion of calcium and a hormonal product called pentagastrin.

Medullary carcinomas of the thyroid

Early diagnosis is essential and asymptomatic family members should be checked regularly. If serum calcitonin levels are elevated or become abnormally elevated following stimulation, the patient's thyroid gland should be removed completely, followed by treatment with replacement doses of thyroxine.

Rickets, osteomalacia, and hypervitaminosis D. Vitamin D deficiency, known as rickets in children and osteomalacia in adults, was a worldwide problem, particularly in temperate zones, until the 1920s when it was found that it could be cured by exposure to light and by the administration of cod liver oil, a substance high in vitamin D. Affected individuals have soft bones, the literal meaning of the term osteomalacia. Their bones become distorted, resulting in bow legs, a bulging forehead, distortion of other bones of the head (craniotabes), and enlargement of the junctions of the ribs with the rib cartilage on the chest (rachitic rosary). These distortions are caused by the generation of excessive amounts of uncalcified bone in an attempt, in effect, to make up for the deficient calcium deposition. Healing takes place promptly with vitamin D supplements, and the disease has become rare with the irradiation of milk and other forms of preventive nutrition.

Vitamin D supplements

Osteomalacia also may be produced in patients suffering from intestinal malabsorption in which ingested vitamin D is not absorbed through the intestinal lining and then into the body. In rare instances families are afflicted with vitamin D-resistant rickets, in which enzymes for the production of the more potent vitamin D metabolites are missing, although this enzyme deficiency can be overcome by administering large doses of vitamin D. Finally, some drugs used to combat seizure disorders (phenytoin and barbiturates) may interfere with the formation of active vitamin D metabolites and thus cause osteomalacia. Because the conversion of 25-hydroxycholecalciferol to the potent derivative of vitamin D_3, 1,25-dihydroxycholecalciferol, takes place primarily in the kidney, this process is impaired in severe kidney disease. (The resulting bone disease, a form of osteomalacia, is known as renal osteodystrophy.)

Ingestion of megadoses of vitamin D produces bone disease associated with hypercalcemia. Treatment, of course, includes the discontinuance of the vitamin D supplements. Reversal of the process can be hastened by the administration of one of the cortisone family of drugs. Occasionally, as in the case of sarcoidosis, there is an abnormal sensitivity to vitamin D or an increased production of vitamin D metabolites, with the absorption of excessive amounts of calcium and the accompanying appearance of hypercalcemia. This disease, too, respond to corticosteroids.

Metabolic bone disease. While the skeleton is usually

thought of as that which is hard and unyielding in the human body, in reality living bone, like many other tissues of the body, undergoes a constant process of breakdown and renewal. This ongoing process of resorption and formation permits the skeleton to adjust to the changes required for healthy functioning, changes ranging from healing fractures to the subtle remodeling necessary to maximize bone strength following alterations in posture or gait. Normal bone provides rigid support, but at the same time it is not brittle. It consists of two major components: (1) a protein matrix consisting mostly of a fibrous protein called collagen; and (2) a mineral portion, mostly complex crystals of calcium and phosphate, which is embedded in the protein component. Bone contains nutritive cells called osteocytes, but the major metabolic activity is carried out by osteoblasts, which generate the protein matrix and osteoclasts (large, multinucleated cells that digest and dissolve bone).

Composition of bone

Only what is called metabolic bone disease, that is, disease which affects all the bones of the skeleton to a lesser or greater extent, is discussed below. These include osteoporosis, osteogenesis imperfecta, osteopetrosis, and Paget's disease of bone. For discussion of such metabolic diseases as rickets, osteomalacia, the bone disease of hyperparathyroidism, and vitamin D intoxication, see above.

The term osteoporosis, taken literally, refers to porous bone. There is simply less bone per unit volume (osteopenia) in osteoporosis. This is true despite the fact that the osteoporotic vertebral body may have collapsed on itself from pressures both from above and from below, forming what is known as a "codfish vertebra." In osteoporosis there is no difficulty with mineralization of bone; rather, the protein matrix is inadequate. There are many reasons for this change. Thinning of bones is part of the process of normal aging, but it can be much accentuated by numerous factors. Most prominent among these are the loss of estrogens in postmenopausal women; multiple forms of nutritional deficiency, including lack of dietary protein; vitamin C deficiency; alcoholism; and low calcium intake.

These deficiencies can occur not only because the required nutrients are not part of the diet but also because of any of a number of disorders associated with poor absorption of nutrients. Osteoporosis occurs rapidly in any person who becomes physically inactive, for example, paralyzed patients or those immobilized by arthritis. It can be produced by drugs such as the corticosteroids, heparin, and anticonvulsants. Estrogen deficiency is an important contributing factor, demonstrated by the fact that bone thinning occurs among those female ballet dancers and long-distance runners in whom menstruation disappears rather than in those in whom it does not. In most of these situations the rate of bone resorption exceeds that of bone formation so that, inevitably, osteopenia occurs.

Osteoporosis in postmenopausal women

Most patients with osteoporosis are women, although the disease does occur in men as well. Of the many causes of osteoporosis, by far the most common is that which occurs in the postmenopausal female. Those who are affected number in the millions, and it is estimated that approximately one-fourth of white women older than 60 years of age have some degree of osteoporosis. Many affected individuals, however, have no symptoms. Others suffer only mild back pain. As the anterior edges of the thoracic vertebra become compressed, the spine bends forward, producing the typical "dowager's hump," with an accompanying loss of height. A compression fracture of a vertebra may be signaled by a sudden, sharp pain in the affected area after minimal or no trauma. It is common, however, for the patient not to recall pain or trauma, and the vertebral fractures may be noted only as incidental X-ray findings. Fractures of the femur after little or no trauma (pathological fractures) are also quite common.

Since estrogens exert a preventive influence on the development of osteoporosis, it occurs most frequently in postmenopausal women. It is not clear why it occurs less frequently in black women than in caucasians. There is evidence that among blacks bone density at maturity is appreciably greater than among whites, so that when bone loss starts, usually several years before the onset of the menopause, those with the greatest bone density are least afflicted. Obesity also exerts a protective effect against osteoporosis, probably because adipose (fatty) tissue is capable of synthesizing estrogens.

Prevention and treatment of osteoporosis

With few exceptions the osteoporotic process (including the osteoporosis of immobilization of the young) is not reversible. The most effective measures are preventive. These include good nutrition and a liberal calcium intake throughout life, but particularly in the early postmenopausal years. Moderate, ongoing physical activity is also essential, but extraordinary long-term exercise (which may result in reduced estrogen secretion) is counterproductive. Estrogen treatment inhibits postmenopausal bone loss at least for the first several years, but whether this is advisable or necessary in all postmenopausal women is not known.

In patients already afflicted with osteoporosis, treatment with calcium, modest doses of vitamin D, or calcitriol may be helpful. Supplemental calcium fluoride also may be helpful, although occasionally at the cost of significant side effects. Again, exercise, even in the frail elderly, is considered an important component and may increase bone density.

Osteogenesis imperfecta, also known as brittle bones, is a rare inherited disease occurring in two forms. In one form, multiple fractures, particularly of the bones of the extremities, occur near the time of birth, and the death rate in afflicted infants is high. The second form is far less severe, with fractures of long bones occurring in adolescence and young adulthood. Associated abnormalities include a blue colour to the whites of the eyes (blue sclerae), along with abnormalities in heart valves. In this disorder there is an inherited defect in the formation of collagen, the protein most abundant in the organic matrix of bone and in heart valve tissue. There is no known treatment.

Osteopetrosis is another rare hereditary disease, characterized by abnormally dense bones that tend to crowd out the bone marrow. The severest form occurs in infants and was uniformly fatal until bone marrow transplantation emerged as a dramatically successful form of treatment.

Paget's disease

In a strict sense, Paget's disease is not a generalized metabolic bone disease; rather, it is a localized disease that may be disseminated to include a large portion of the skeleton. For this reason, it can be included with the metabolic disorders of bone.

The most graphic and detailed description of this disorder was provided by Sir James Paget, a prominent English surgeon. Paget believed the disease resulted from inflammation, and for this reason he called it osteitis deformans. This notion was soon discredited and many other possible causes were considered more seriously, but it now appears that Paget was correct. Under the ultramicroscope, structures that very closely resemble viruses have been seen in the osteoclasts of patients suffering from Paget's disease. The osteoclasts are extraordinarily active, digesting bone at a very rapid rate and at the same time activating a "coupling factor" that leads to a compensatory increase in bone synthesis by local osteoblasts. The result is a changed, "chaotic" bone structure leading to bone weakening and deformities.

The patient with classical, advanced Paget's disease has a large skull, a shortened spine, and bowed thighs and legs, producing a simian appearance. Pathological fractures are common, and the patient's course may be threatened by complications such as impingement of distorted vertebrae on the spinal cord, which threatens paralysis. Occasionally, the pathological stimulation of bone turnover leads to a transformation into bone cancer. There is no known cure, but the process can be suppressed effectively with a number of therapeutic agents, including salmon calcitonin, one of the diphosphonates, or plicomycin (mithramycin).

Fibrous dysplasia

Fibrous dysplasia also is a disseminated, rather than generalized, bone disease, and its cause is unknown. It may be monostotic (localized to one bone) or polyostotic. The disease leads to a gross distortion of bone structure that may result in a grotesque appearance of facial features. There are often accompanying patches of tan pigmentation (café au lait spots) and if the base of the skull is involved, particularly in females, puberty may occur at an inordinately young age (precocious puberty). (T.B.S.)

THE PANCREAS

The discovery of insulin in 1921 by a Canadian surgeon, Frederick Banting, with the assistance of a medical student, Charles Best, was one of the most dramatic events in modern medicine. It not only saved the lives of innumerable patients affected with childhood diabetes but it also ushered in present-day understanding of the complexities of the endocrine pancreas. The importance of the endocrine pancreas lies in the fact that its principal hormone, insulin, plays a central role in the regulation of energy metabolism and that a relative or absolute deficiency of insulin leads to diabetes mellitus, still a leading cause of disease and death throughout the world.

Anatomy. In humans the pancreas weighs approximately 80 grams, has roughly the configuration of an inverted smoker's pipe, and is situated in the upper abdomen. The head of the pancreas (equivalent to the bowl of the pipe) is immediately adjacent to the duodenum, while its body and tail extend across the midline nearly to the spleen. The bulk of pancreatic tissue is devoted to its exocrine function, the elaboration of digestive enzymes that are secreted via the pancreatic ducts into the duodenum.

The islets of Langerhans

The endocrine pancreas consists of the islets of Langerhans. Approximately 1,500,000 islets, weighing about one gram in total, are scattered throughout the gland. The embryonic origin of the cells that make up the islets is not clear; both endodermal and neuroectodermal precursors have been proposed. Approximately 75 percent of the cells in each islet are the insulin-secreting beta (B) cells, which tend to cluster centrally (Figure 13, top). Around the periphery lie the alpha (A), delta (D), and F (or PP) cells, which secrete glucagon, somatostatin, and pancreatic polypeptide, respectively. Each islet is supplied by one or two minute arteries that branch into numerous capillaries; from this network, capillaries emerge to coalesce into small veins outside the islet. The islets also are richly supplied with autonomic nerves. Thus, islet function may be modulated by neural control, by circulating metabolites and hormones, and by secretion of hormones locally (paracrine effects).

The principal function of the endocrine pancreas is the secretion of insulin and other polypeptide hormones necessary for the orderly cellular storage and retrieval of such dietary nutrients as glucose, amino acids, and triglycerides.

Hormones. *Insulin.* Insulin, produced by the beta cells of the islets of Langerhans, is a moderate-sized protein composed of two chains, the alpha chain (with 21 amino acids) and the beta chain (with 30), linked by sulfur atoms. Insulin is derived from a larger prohormone molecule

Proinsulin

called proinsulin. Proinsulin is relatively inactive, and normally little of it is secreted. It contains a connecting peptide, or C-peptide, composed of 31 amino acids with an additional amino acid at either end linked to the alpha and beta chains, respectively. As is the case with other prohormones, the connecting peptide of proinsulin is cleaved off before insulin is released into the circulation. Insulin leaves the pancreas through veins, which empty into the portal vein perfusing the liver. Typically the pancreas of a normal adult contains approximately 200 units (eight milligrams) of insulin; the average daily secretion of insulin into the circulation ranges between 35 and 50 units.

Although a number of physiological events influence insulin secretion, the most important is the concentration of glucose in the arterial (oxygenated) blood perfusing the pancreas. When the plasma glucose level rises, insulin release is stimulated; as plasma glucose falls, so does the rate of insulin secretion. Even during prolonged fasting, however, a baseline secretion of insulin continues. Insulin secretion also is influenced by neurotransmitters interacting with islet cell receptors, particularly those that bind norepinephrine.

The action of insulin can be appreciated by considering its effect on three tissues important in metabolism (adipose tissue, muscle, and liver) and by noting the consequences of its deficiency in diabetes mellitus (see below *Diabetes mellitus*). Insulin has profound effects on adipose tissue and lipid metabolism: it permits the entry of glucose into the fat cell (adipocyte) and then stimulates the metabolism

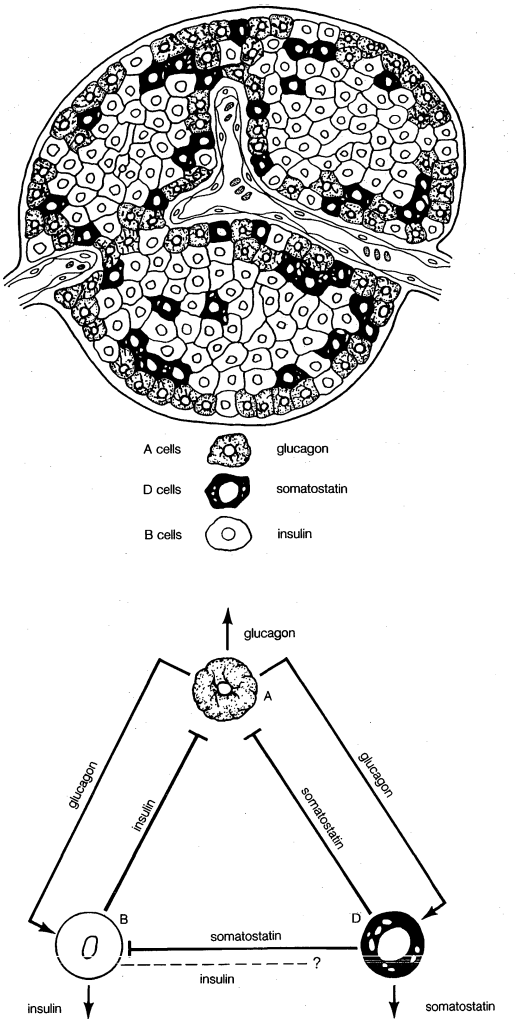


Figure 13: *Relationships of the cells of the islets of Langerhans.*
(Top) Cross-section of a normal islet. (Bottom) Possible paracrine cell–cell interactions within the islets. Stimulatory actions of the peptides on the neighbouring cells are represented by long arrows; inhibitory actions are represented by short bars. Endocrine secretion into the circulation is shown by short arrows.
From R.H. Unger and L. Orci, *New England Journal of Medicine*, vol. 304, p. 1518 (1981); reprinted by permission of the *New England Journal of Medicine*

of glucose once it is in the cell. The presence of glucose within the adipocyte, in turn, leads to increased formation of fatty acids and triglycerides. Insulin has a stimulatory effect on lipoprotein lipase, an enzyme located in the walls of the capillaries in adipose tissue and one that is required for splitting circulating triglycerides, a necessary step before the fatty acid contained in triglycerides can enter fat cells. Finally, insulin is the most potent inhibitor of the release of stored fatty acids. As the level of plasma insulin rises, the release of fatty acid, or lipolysis, is markedly suppressed; conversely, as insulin falls, the release of fatty acid accelerates.

Insulin stimulates the transport of glucose and amino acids into muscle cells and prompts the conversion of amino acids into protein. Thus, insulin is required to replete the glycogen, a stored form of glucose, that is oxidized during exercise and to replenish protein needed for muscle growth and repair. Insulin is not required for the transport of glucose into liver cells, but the hormone profoundly affects intracellular metabolism in the liver. It promotes glycogen formation, stimulates the utilization of glucose, and suppresses those enzymes necessary for new glucose formation (gluconeogenesis) and glycogen breakdown (glycogenolysis). The overall effect of insulin is to increase glucose utilization and storage and decrease its release by the liver.

Overall effect of insulin

Glucagon. Glucagon is produced by the alpha cells

of the pancreas and also is secreted by cells scattered throughout the gastrointestinal tract. A number of forms of glucagon have been found; the biologically active one appears to contain 29 amino acids. Radioimmunoassays can distinguish between pancreatic glucagon and similar peptides from the gut. Circulating glucagon levels are high in the fasting state. Secretion is stimulated by amino acids and gastrointestinal peptide hormones. Normally, ingested glucose is a potent suppressor of glucagon release, an effect that is probably mediated by an increase in circulating insulin. Secretion of glucagon also is inhibited by free fatty acids and by somatostatin and appears to be modulated by the autonomic nervous system. Circulating glucagon binds to specific receptors on the surface of liver cells (hepatocytes), leading to the breakdown of liver glycogen into glucose, which is then released into the blood. Glucagon is estimated to be responsible for most of the hepatic glucose production after an overnight fast.

Somatostatin. Somatostatin, a peptide that was discovered initially in the hypothalamus (see above *Hypothalamus: Growth hormone-releasing hormone*), contains 14 amino acids, is produced by the D cells of the islets, and has a number of effects on digestion. It inhibits gastrointestinal motility and blood flow, secretion of stomach acid, secretion of pancreatic exocrine, and the absorption of triglyceride from the bowel.

Effects of pancreatic hormones

In summary, it appears that the hormones insulin, glucagon, and somatostatin act in concert to control the flow of nutrients into and out of the circulation. The relative concentrations of these hormones regulate the rates of absorption and peripheral disposal of substances such as glucose, amino acids, and fatty acids. The anatomic proximity of the B, A, and D cells in the islets is significant. Somatostatin and glucagon appear to have a paracrine relationship whereby they influence the secretion of each other, and both affect the rate of insulin release (Figure 13, bottom).

Pancreatic polypeptide. Pancreatic polypeptide, secreted by the F (or PP) cells, contains 36 amino acids. Circulating levels rise following ingestion of a meal. An increase in the level of free fatty acids in the blood suppresses its secretion. Pancreatic polypeptide can inhibit gallbladder contraction and pancreatic exocrine secretion, but its biologic role is uncertain.

Hormonal control of energy metabolism. The functions of the pancreatic hormones, particularly insulin and glucagon, can best be appreciated by considering their roles in maintaining glucose homeostasis and in regulating nutrient storage. An adequate supply of glucose is required for optimal body growth and development and for the health of the central nervous system, for which glucose is the major (and usually the only) source of energy. It is not surprising, therefore, that elaborate mechanisms have evolved to ensure that a normal plasma glucose level is maintained regardless of whether a person is feasting or fasting. Another requirement for survival is the ability of the body to store excess nutrient fuel for recall and use during periods of scarcity. Adipose tissue serves this need. Compared to carbohydrate and protein, fat yields twice the calories per gram. Furthermore, adipose tissue contains less than 10 percent water. Thus, a kilogram of fat has 10 times the caloric value of a portion of muscle of the same weight. Following the ingestion of a meal, the carbohydrate content is assimilated as glucose, leading to an elevation in blood glucose and to an increase in plasma insulin from 10 microunits per millilitre to 50–100 microunits per millilitre. This high level of insulin promotes glucose uptake by the liver, adipose tissue, and muscle. Fatty acids and amino acids derived from the digestion of fat and protein are also deposited in the liver and peripheral tissues. Glucose production by the liver is inhibited, and brain metabolism is fueled by dietary glucose. Insulin also suppresses lipolysis so that the concentration of free fatty acids in the plasma falls. Thus, the "fed," or anabolic, state is characterized by nutrient storage dependent on an increase in circulating insulin.

The anabolic state

Within a few hours after a meal, when gastrointestinal absorption of nutrients is complete, the level of insulin falls and hepatic production of glucose resumes, sustaining the needs of the brain. Similarly, lipolysis increases, providing fuel for muscle. After a longer period of fasting, (*e.g.,* 12 to 14 hours or overnight), the insulin level falls still lower and plasma glucagon increases. Hepatic glycogen becomes depleted, and glucose production is achieved by gluconeogenesis, a process requiring precursor carbon molecules such as the amino acid alanine from muscle breakdown and glycerol from lipolysis. Thus, the "fasting," or catabolic, state is characterized by a low level of insulin, an increased concentration of glucagon, and a withdrawal of stored nutrients.

With further fasting, lipolysis continues to increase for a few days before it plateaus at a high rate. A large proportion of elevated fatty acids are converted to the "ketone bodies" in the liver, a process enhanced by the high level of glucagon. The brain, previously an avid and fastidious consumer of glucose, begins to use ketones as well as glucose. Eventually, more than one-half of the brain's daily metabolic energy needs are met by the ketone bodies, thus substantially diminishing the need for hepatic glucose production. The decrease in gluconeogenesis reduces the need for protein-derived amino acids, sparing muscle and making survival during prolonged fasting possible. Starvation is characterized by very low levels of insulin, elevated concentrations of glucagon, and very high concentration of circulating free fatty acids and ketones.

Insulin actions

In summary, in the fed state insulin mediates (1) the transport of glucose into body tissues (to be consumed as fuel or stored as glycogen), (2) the transport of amino acids into tissues (to build or replace protein), and (3) the transport of glucose and fatty acids into adipose tissue (to provide a fuel depot for future energy needs). During fasting, insulin levels are depressed and the opposite sequence of events occurs, modulated by glucagon and other "anti-insulin" hormones such as cortisol from the adrenal cortex.

Diseases and disorders. *Diabetes mellitus.* A relative or absolute deficiency of insulin results in the disease diabetes mellitus, by far the most common disorder of the endocrine system. The number of individuals with diabetes doubles every 15 years. While insulin, discovered by Banting and Best in 1921, can prevent early death from diabetic coma, insulin treatment does not prevent the chronic, disabling complications of the disease. Statisticians list diabetes mellitus among the top 10 causes of death in the United States, for example, and cite it as the leading cause of blindness and uremia.

Types of diabetes

In the United States, the National Institutes of Health has classified diabetes into a number of types. Type I, or insulin-dependent diabetes mellitus (IDDM), formerly termed juvenile-onset diabetes, can occur at any age of life. Affected individuals have insulin deficiency due to islet cell loss and may become comatose when exogenous insulin is withheld. Type II, or non-insulin-dependent diabetes mellitus (NIDDM), previously called maturity-onset diabetes, also can occur at any age but is most common in adults. Affected individuals are not prone to coma except in the presence of stress, although they often require insulin to control hyperglycemia. The majority are obese. Other types are a miscellaneous group, formerly called secondary diabetes, and include diseases attacking the pancreas (*e.g.,* hemochromatosis, pancreatitis), and syndromes characterized by insulin antagonism (*e.g.,* Cushing's disease, acromegaly). The term impaired glucose tolerance (IGT) is applied to those who have oral glucose tolerance tests (OGTT) that exceed normal levels but are not sufficiently abnormal to justify the diagnosis of diabetes mellitus. Most of these individuals do not progress to overt diabetes and do not develop the chronic complications of the disease. The term gestational diabetes mellitus (GDM) is reserved for diabetes or glucose intolerance that develops, or is first recognized, during pregnancy. Patients usually revert to normal glucose tolerance following pregnancy.

In order to diagnose diabetes mellitus in an apparently healthy adult, a physician must observe either two fasting plasma glucose values greater than 140 milligrams per decilitre or any two values greater than 200 milligrams per decilitre following a 75-gram oral glucose load. Criteria

for the diagnosis of glucose intolerance include a fasting plasma glucose value between 115 and 140 milligrams per decilitre, a two-hour postprandial value between 140 and 200 milligrams per decilitre, and at least one value greater than 200 milligrams per decilitre during a standard oral glucose tolerance test.

Causes of diabetes

It seems likely that there are two distinct causes for IDDM and NIDDM. A genetic factor appears to be more important in NIDDM, since analysis of a large series of identical twins has shown a concordance (the appearance of the trait) in both twins of more than 90 percent for NIDDM, while in IDDM the rate is about 50 percent. This relatively low incidence of the disease among the identical twins of insulin-dependent diabetics suggests that other factors are important. One such factor may be immune-related. Among insulin-dependent diabetics, there is a relatively high prevalence of certain patterns of the inherited tissue compatibility antigens (HLA), while in NIDDM the prevalence of these HLA types is normal. In addition, there is a high prevalence of autoantibodies to islet cells found in the sera of insulin-dependent diabetics, along with inflammation of the islets. There is evidence that in some cases of IDDM, viral infections may play a role. Coxsackie virus B_4 has been isolated from the pancreas of a child who died accidentally shortly after the onset of diabetes. This virus was cultured and found to cause beta cell damage when injected into mice. Further evidence for an infectious factor in the causation of type I diabetes mellitus is the seasonal appearance of new cases of the disease.

Viral and autoimmune factors

It may be that viral and autoimmune factors combine to cause diabetes, the viral infection of the pancreas leading to the release of proteins into the circulation that are recognized as foreign by the victim's immune system. In this theory, autoantibodies cause destruction of the beta cells of the pancreas. This thesis suggests the possibility of preventing the disease either by immunization against suspect viruses or early treatment with immunosuppressive drugs. Patients with NIDDM appear to suffer from resistance to insulin along with abnormal secretion of the hormone. In fact, these patients may initially have higher than normal concentrations of plasma insulin. The primary site of resistance is likely to be within the cell (*i.e.*, a postreceptor defect), although a receptor abnormality may also be implicated. In many, but not all type II patients, resistance to insulin is linked to obesity.

A relative or absolute deficiency of insulin results in hyperglycemia, the central biochemical feature of the disease. Hyperglycemia ensues because of impaired transport of glucose into muscle and adipose tissue and the increased release of glucose into the circulation by the liver. Above a glucose concentration of about 180 milligrams per decilitre the kidney tubules are unable to reabsorb all of the glucose filtered by the glomeruli. The excretion of glucose by the kidney requires a simultaneous movement of water out of the plasma and into the kidney, from which it is excreted. The subsequent increase in the relative concentration of solutes in the water-depleted plasma in turn stimulates the thirst centres of the hypothalamus in the brain. Three classic conditions thus result: polyuria (excretion of a large volume of urine in a specific amount of time); polydipsia (excessive, long-term thirst); and polyphagia (voracious eating). All can be explained in terms of the body's loss of large quantities of glucose and water, which results in a compensatory increase in hunger and thirst. With more severe insulin deficiency, hyperglycemia and glycosuria intensify, liquid intake falls behind urinary loss, and dehydration and shock ensue. The rate of fatty acid release from adipose tissue is greatly accelerated. Much of the fatty acid reaching the liver is converted to the ketone bodies acetoacetic acid and beta-hydroxybutyric acid. Both substances lower the pH of the blood, normally held at a pH of 7.4. As the acidic state progresses, there is depression of cerebral and myocardial function, culminating in coma and death. Appropriate fluid therapy and administration of insulin is life saving. The blood sugar is lowered, dehydration and shock are reversed, and the blood pH is restored to normal.

Prolonged survival of patients with diabetes mellitus has led to an increasing incidence of chronic complications, most of which can be explained by changes in the patient's blood vessels. Small blood vessel disease (microangiopathy) is unique to diabetes; the principal feature seen under the microscope is thickening of the walls of the capillaries. With time, affected capillaries become leaky, leading to changes in the retina (retinopathy) and kidney (nephropathy). Ultimately, there may be retinal hemorrhage leading to blindness and severe impairment of renal function causing uremia. Diabetic patients are also afflicted by an increased incidence of large vessel disease. Microscopically, hardening of the arteries (atherosclerosis) in the diabetic is not different from that seen in nondiabetic individuals; however, it occurs earlier and progresses faster in diabetic patients. Premature coronary artery disease is a common cause of death among diabetics. The large arteries of the lower extremities are often affected, contributing to the high incidence of foot ulceration and gangrene and resulting in amputation.

Complications of diabetes

Not all complications are directly related to vascular disease. Early cataract formation, impaired function of the autonomic nervous system, and peripheral nerve damage (neuropathy) cannot be fully explained by blood vessel changes. Autonomic nervous system dysfunction may be manifested by gastric retention, chronic diarrhea, incomplete emptying of the bladder, impotence, and low blood pressure when standing. Diabetic neuropathy often affects the lower extremities, causing either loss of feeling or disagreeable sensations of burning or itching. It is not known if these complications are due to long-standing hyperglycemia and insulin deficiency or are caused, at least in part, by an unidentified factor.

Treatment of diabetes

There are three basic components in the treatment of IDDM: insulin, diet, and exercise. Insulin is prepared in a number of forms, providing short, intermediate, and long actions to accommodate the specific needs of individual patients. Sources of insulin traditionally have been from pork and beef pancreases, but insulin with a structure identical to that produced by human islets is now widely available. It is either synthesized using recombinant DNA technology or prepared by the chemical alteration of porcine insulin. Insulin is given by one or more injections each day or by continuous infusion using an insulin pump, a computerized device the size of a deck of cards, which is worn by the patient and delivers a preset amount of hormone throughout the day. A typical diabetic diet contains sufficient total calories to maintain ideal body weight and consists of carbohydrate, fat, and protein and of ample fibre. Simple sugars and alcohol are prohibited. Compared to insulin and diet, exercise is more difficult to measure. Ideally, the diabetic exercises a fixed amount each day, and insulin and diet are tailored to accommodate that amount.

The goals of treatment include maintenance of the blood sugar near or within normal limits, freedom from hypoglycemia, and an acceptable life-style. The blood sugar usually can be monitored at home by the patient. Measurement of that portion of hemoglobin complexed to glucose provides another index of the adequacy of blood sugar control. Proteins exposed to glucose-containing solutions undergo glycosylation (*i.e.*, a certain number of glucose molecules become irreversibly fixed to the protein molecule). The higher the glucose concentration, the greater the degree of glycosylation. In normal humans approximately 6 percent of circulating hemoglobin is glycosylated; in poorly controlled diabetics, the figure may be 14 percent or more. Glycosylation of structural proteins, such as those in the basement membranes of capillaries, may play a role in the pathogenesis of the chronic complications of diabetes.

In the treatment of NIDDM, diet and exercise again are important. In the obese patient, evidence of diabetes mellitus may disappear if the patient can achieve and maintain ideal body weight. Insulin or a blood-sugar-lowering drug, however, may be required to control blood sugar.

Hypoglycemia. Although a wide array of human ills are attributed to low blood sugar (hypoglycemia), well-documented hypoglycemia is not common except among insulin-dependent diabetics. Regardless of the underlying

cause, the manifestations of hypoglycemia evolve in a characteristic pattern. Mild hypoglycemia causes hunger, fatigue, tremor, perspiration, weakness, and anxiety. These same symptoms often appear in a variety of conditions other than hypoglycemia, however. To implicate hypoglycemia properly, such symptoms should be associated with a blood glucose of less than 40 milligrams per decilitre and be promptly relieved by the administration of glucose. More severe hypoglycemia leads to blurred vision, impaired mentation, and bizarre behaviour. A staggering gait and irrational, hostile behaviour are frequently misinterpreted as drunkenness. Finally, the patient becomes comatose and may develop generalized seizures. If severe hypoglycemia remains untreated, permanent brain damage or death can result.

The principal causes of hypoglycemia can be grouped into two large categories: fasting and fed (or "reactive"). The time of day at which hypoglycemia occurs provides a clue to the underlying cause. Since liver production sustains the blood glucose level during periods of fasting, rare, inherited disorders that cause impairment in glycogen storage or gluconeogenesis lead to fasting hypoglycemia. This typically occurs in the early morning hours after eight or nine hours of fasting. Fasting hypoglycemia is also caused by insulin-secreting islet cell adenomas. In these cases, hypoglycemia is prevented during the waking hours by frequent eating, leading to weight gain.

Reactive hypoglycemias

Far more common are the reactive hypoglycemias, triggered by the assimilation of glucose following a meal. Normally, the secretion of insulin is commensurate with the degree of postprandial elevation of the blood sugar. After surgery that impairs the reservoir function of the stomach, ingested glucose is dumped into the duodenum and upper jejunum, where it is rapidly absorbed. The resulting excessive hyperglycemia stimulates a brisk release of insulin, leading to moderately severe hypoglycemia within two hours of the meal. There is another group of patients who assimilate glucose at an excessive rate even though they have not had gastrointestinal surgery. Such individuals typically have symptoms three to four hours after the ingestion of a large quantity of glucose. Manifestations of hypoglycemia usually can be avoided in reactive hypoglycemia by restricting the amount of glucose in the diet. The most frequent cause of clinically significant hypoglycemia is self-administration of insulin by the diabetic patient. This may result from an excessive dose of insulin, inadequate dietary amounts, or excessive physical activity. Rarely, hypoglycemia may be self-induced by emotionally disturbed patients with access to insulin.

Tumours of the pancreas. Inappropriate hypersecretion of pancreatic hormones may be due to diffuse hyperplasia (abnormal multiplication) of the secretory cells, adenomas (benign tumours), or carcinomas (malignant tumours). Hypersecretion of insulin is most frequently due to a single insulin-producing adenoma. Single or multiple insulinomas may occur as part of the syndrome of multiple endocrine neoplasia. Malignant insulinomas are less common. Diffuse hyperplasia of beta cells (nesidioblastosis) may cause hypoglycemia in infants. Glucagon-secreting tumours (glucagonomas) produce the "diabetes-dermatitis syndrome." Patients have mild diabetes, anemia, and a red, blistering rash that appears in one area of the body and then fades, only to reappear at a different site. Patients have elevated plasma glucagon levels, but marked hyperglycemia is prevented by an offsetting increase in insulin secretion.

Somatostatin-producing tumours are difficult to diagnose because findings are nonspecific and include diabetes mellitus, gallstones, excessive fat in the stool, indigestion, and diminished secretion of gastric acid. Plasma somatostatin levels are increased when measured by radioimmunoassay, and both insulin and glucagon concentrations are decreased. Pancreatic polypeptide-secreting islet cell tumours have been found in patients with the syndrome of multiple endocrine neoplasia. Pancreatic tumours may also be the source of "ectopic" hormone secretions (in which a hormone is secreted from a tissue type that normally does not secrete it; see below *Ectopic hormone and polyglandular disorders*). (T.W.B.)

THE ADRENAL CORTEX

Anatomy. The adrenal glands lie on the upper inner surface of each kidney. Each gland consists of two parts that are quite distinct anatomically, embryologically, and functionally. The inner core (adrenal medulla) is discussed separately below. The outer covering (adrenal cortex) is derived from the fetal mesodermal ridge, a structure that also gives rise to the kidneys so that the juxtaposition of the two organs is not surprising. Within the adrenal cortex are three zones known as the outer (zona glomerulosa), the middle (zona fasciculata), and the inner (zona reticularis). Under the microscope the cells are rather typical endocrine cells; the distinction between zones is made by differing staining characteristics.

The zones of the cortex

Hormones. Adrenocortical cells synthesize and secrete chemical derivatives (steroids) from cholesterol, the major animal sterol. While cholesterol can be synthesized in many body tissues, further differentiation into steroid hormones takes place only in the adrenal cortex and in its embryological cousins, the ovaries and the testes.

The adrenal cortex is capable of synthesizing all of the steroid hormones produced by the body, including the progestogens and estrogens (see below *The ovary*), androgens (see below *The testis*), mineralocorticoids (which are secreted from the zona glomerulosa), and glucocorticoids (which are synthesized and released from the zona fasciculata and zona reticularis of the adrenal cortex). Although upwards of 60 steroids are manufactured in the adrenal cortex, only a few members of these three major categories are important in body functioning.

Aldosterone. The biologic effect of aldosterone, the principal mineralocorticoid produced by the zona glomerulosa, is to set in motion a set of reactions at the cell surface of all body tissues in order to enhance the uptake and retention of sodium in all cells and the extrusion of potassium from them. Such fluxes of sodium and potassium following the administration of aldosterone are detectable even in glandular secretions, such as sweat. It also has a major impact on kidney function, acting on the renal tubules to retain sodium within the circulation while increasing the excretion of potassium into the urine. At the same time, by increasing the reabsorption of bicarbonate by the kidney, aldosterone tends to decrease the acidity of body fluids.

Cortisol. Cortisol (hydrocortisone) is the major human glucocorticoid. It exerts multiple and varied effects. It also serves as a mineralocorticoid but is considerably less effective than aldosterone. Cortisol plays a major role in the body's response to stress. In fasting, for example, it sustains the blood sugar concentration by blocking the egress of glucose into all tissues other than the critically important brain and spinal cord, while it simultaneously increases the breakdown of protein from muscle and other organs and hastens the conversion of newly generated amino acids to glucose to replenish the supply constantly being consumed by the brain.

"Permissive" actions

In addition, glucocorticoids have a "permissive" action for many chemical reactions in the body; that is, their presence is necessary for the action to occur, but they themselves do not initiate it. For example, the secretion of acid into the stomach does not occur in the total absence of glucocorticoids, but, in the presence of normal amounts of cortisol, acid can be excreted in small or large amounts as the body requires.

Cortisol, along with more potent and longer-acting synthetic derivatives like prednisone, methylprednisolone, and dexamethasone exerts powerful anti-inflammatory effects. Physicians take advantage of these properties in treating patients with serious inflammatory illnesses such as rheumatoid arthritis, disseminated lupus erythematosus, and multiple sclerosis. If, however, the inflammation has a bacterial or viral origin, the steroids may do more harm than good because the spread of the infection is facilitated while the signs of inflammation are masked (see IMMUNITY). Finally, corticosteroids in large doses impair the functioning of the immune system so that the production of harmful antibodies, such as those produced in allergic diseases, may be suppressed. It is important to note that these beneficial effects are offset by serious side effects

of large-dose, long-term corticosteroid therapy, effects that closely mimic many of the symptoms of Cushing's syndrome (see below).

Adrenal androgens. Ordinarily, adrenal estrogens do not play an important role in the body's economy, but adrenal androgens do make a significant contribution. These androgens are not as potent as testosterone, the major steroid secreted by the testis, but a number of them, including androstenedione, dehydroepiandrosterone (DHEA), and its sulfate (DHEAS) may be converted to stronger androgens such as testosterone. Although little androgen is secreted before puberty, the output increases dramatically at puberty so that the adrenal cortex makes a significant contribution, known as the adrenarche, to developmental changes in both sexes.

All steroid hormones, including those from the adrenal cortex, are bound to steroid-binding globulins (transcortin) in the circulation and are released at the surface of a target cell. The steroid passes into the cell cytoplasm and is bound to an intracellular binding protein and thence is transported into the cell nucleus. There the hormone exerts its effect by modulating gene activity so that the synthesis of some proteins is stimulated while that of others might be inhibited. The net effect is the biologic action noted at the physiological or pathological level. The steroid hormones undergo inactivation in a complex series of transformations principally in the liver but in other tissues as well, leading to a total loss of hormonal activity.

Regulation of hormone secretion. The three classes of corticosteroids (the mineralocorticoids, the glucocorticoids, and the adrenal androgens) are regulated largely by separate mechanisms. Glucocorticoids are regulated by way of the classical hypothalamic-hypophyseal feedback system shown in Figure 3. Within the family of glucocorticoids, the cortisol level is the one most closely guarded. Furthermore, the ongoing feedback control is modulated by hypothalamic biorhythmic activity illustrated in the case of cortisol in Figure 9. When the individual is exposed to physical or emotional stress, the self-regulating mechanism is interrupted and plasma cortisol is increased to deal with the stress. Adrenal androgen secretion is controlled primarily by ACTH, although there is evidence that prolactin stimulates the secretion of adrenal androgens as well.

Regulation of corticosteroids

Aldosterone secretion is modulated directly by serum electrolyte levels. Lowered serum sodium concentrations enhance aldosterone secretion, but a far more potent stimulus is a high serum potassium level.

A major regulator of aldosterone secretion is the renin-angiotensin system, although ACTH also stimulates mineralocorticoid secretion. Renin is an enzyme secreted into the blood plasma from specialized cells encircling the arteriole located at the entrance to the glomerulus (the renal capillary network that is the filtration unit of the kidney). Renin secretion is inhibited when these cells, contained in what is known as the juxtaglomerular apparatus, are compressed by dilatation of this entering (afferent) arteriole provoked by an increase in plasma volume. When plasma volume decreases, renin secretion is stimulated.

Renin catalyzes the conversion of a plasma protein, angiotensinogen, into an active decapeptide, angiotensin. Angiotensin is a potent stimulator of arteriole constriction and aldosterone secretion. Both actions result in higher blood pressure, the first by increasing resistance to the flow of blood ejected by the heart, and the second by increasing total plasma volume. These are key responses when blood pressure falls to a dangerously low level. On the other hand, excessive renin secretion can lead to ongoing high blood pressure with its dangerous consequences to the health of the patient.

For a number of years investigators have sought a factor secreted by the hypothalamus that would specifically modulate aldosterone secretion in the same negative feedback fashion that relates pituitary corticotropin to the adrenocortical glucocorticoid, cortisol. Several candidates, including melanotropin (MSH) and endorphins, have emerged. Whether either or both of these hormones completely fulfill this role remains uncertain.

More recently, a new group of factors, the atrial natriuretic (sodium-excreting) peptides (atriopeptin, atrin), have been characterized. These hormones are secreted into the blood when the upper chambers of the heart, the atria, are stretched by an expanded volume of blood. The major polypeptide isolated from human atria, atrin, contains 28 amino acids. In general, the actions of atrin oppose those of angiotensin; atrin blocks the contractions of muscles in the walls of arteries so that the arteries dilate, and atrin inhibits the synthesis and secretion of aldosterone. Furthermore, it inhibits the release of renin from the juxtaglomerular cells, and finally it acts directly on the kidney to increase the excretion not only of urine but also the sodium chloride, potassium, magnesium, and phosphorus contained in it. This powerful natriuretic action may have important therapeutic applications in patients with heart failure, high blood pressure, liver disease, or other illnesses associated with the retention of fluid. Finally, it should be noted that the adrenal glands are influenced not only by endocrine factors but also by neural influences. The neurotransmitter dopamine is a powerful suppressor of aldosterone secretion, while serotonin may have a stimulating effect.

Atrial natriuretic peptides

Diseases and disorders. *Adrenal insufficiency (Addison's disease).* Adrenal insufficiency is a rare disease. In the past it was caused most commonly by destruction of both adrenals in tuberculosis patients. More recently, it has been found that destructive autoantibodies are most often the cause, sometimes as part of the inherited syndrome of multiple endocrine deficiencies (see below *Ectopic hormones and polyglandular disorders*). Other infectious diseases such as histoplasmosis may also destroy both adrenals. The adrenal glands may be involved in many other pathological processes (for instance, invasion by cancer), but adrenal insufficiency does not supervene because more than 90 percent of the total of adrenal cortical tissue must be destroyed before it becomes incapable of providing for the body's needs. Adrenal insufficiency may be secondary to diseases of the pituitary or hypothalamus, resulting in deficiencies of corticotropin or CRH, respectively.

Addison's disease, if undiagnosed, leads to death. The onset of Addison's disease is often gradual and puzzling to both patient and physician. There is an increasingly generalized weakness along with an inordinate tiredness after physical activity. The patient loses appetite and weight and suffers occasional bouts of vomiting and diarrhea. There is increasing pigmentation, not only in exposed areas but also in the nails and in the skin creases. The patient's blood pressure falls, and there may be episodes of fainting upon arising from bed or from a chair. The patient must eat regularly because even minor delays result in hypoglycemic episodes. If the disease is caused by an infectious agent, there may be calcification in the area of the adrenals seen on X-ray examination of the abdomen.

Addison's disease

The symptoms intensify over a period of months until, either spontaneously or as the result of physical stress, such as trauma or an intercurrent illness, the patient suffers acute adrenal insufficiency, known as Addisonian crisis, and experiences a catastrophic change in status. With intensified vomiting, diarrhea, and fever and with a precipitous fall in blood pressure, the patient goes into shock and dies.

Addisonian crisis may occur also in individuals who have no previous adrenal disease. During or shortly after birth some infants suffer bilateral massive adrenal hemorrhage. A similarly destructive hemorrhage can occur in adults, especially those who are treated with anticoagulants like heparin and undergo an operation or other trauma. Cortisol given intravenously is life-saving.

In chronic adrenal deficiency the patient can be kept alive and well with modest doses of cortisol taken orally, often along with a synthetic mineralocorticoid. Occasionally, salt tablet supplements are useful. The dosage must be sharply increased during periods of acute illness or injury. Before the advent of these simple therapeutic measures patients with Addison's disease died within two to three years after diagnosis. Such patients can now look forward to a full life span as long as they are prepared to increase the dosage of cortisol in the event of serious physical stress.

Hypercorticism (Cushing's syndrome). Hypercorticism, the illness resulting from overactivity of the adrenal cortex, exemplifies nicely the medical term syndrome, a constellation of symptoms and signs that together makes up a specific, easily recognized clinical entity, but which has diverse causes. In 1932, the American Harvey Cushing, a pioneer in the field of neurosurgery, described the clinical picture of patients harbouring a specific type of pituitary tumour, an entity that became known as Cushing's disease.

Further studies over the years have revealed that, with minor variations, the clinical picture described by Cushing could also result from at least four other causes: a benign tumour or a cancer of the adrenal cortex; a corticotropin-releasing, hormone-producing hamartoma of the hypothalamus; a number of tumours, both benign and malignant, that ordinarily do not secrete hormones (ectopic hormones that then produce tumours); and finally, the therapeutic administration of large doses of adrenocortical hormones (iatrogenic, or physician-induced, Cushing's syndrome). Thus, ordinarily, the clinician first makes the diagnosis of Cushing's syndrome and then explores further to determine the specific cause so that appropriate, specific treatment can be administered.

Causes of Cushing's syndrome

Cushing's disease results from a hyperfunctioning, corticotropin-producing, benign (rarely malignant) tumour of pituitary corticotrophs. Secreted along with corticotropin is melanotropin (MSH) so that the patient becomes progressively pigmented in a fashion similar to what is seen in patients with Addison's disease. For reasons poorly understood, the patient gains weight in a peculiar distribution; the obesity is confined to the central body areas—the abdomen and back and buttocks—with rather thin extremities. Excess fat deposits occur at both temples, giving rise to a "moon face," and fat may be deposited in the anterior neck ("dewlap"), below the neck posteriorly ("buffalo hump"), around the heart, and even in the spinal canal. Most of the symptoms result from the powerful protein catabolic and gluconeogenetic effects of the glucocorticoids. All patients show progressive weakness and muscle wasting. The skin becomes thin and fragile so that hemorrhages beneath the skin occur frequently. The bone becomes osteoporotic. The increased glucose production may lead to diabetes mellitus as an additional complication. Finally, in women, ovulation is suppressed and there is often amenorrhea along with hirsutism (hairiness), the result of increased adrenal androgen secretion.

Treatment

Treatment is directed against the specific cause. Pituitary tumours are removed surgically, and recurrences may be treated with X-ray therapy. Adrenal tumours also are removed surgically. Adrenocortical carcinomas usually carry a grave prognosis; initially they are treated by surgical removal unless they have already metastasized, in which case a number of drugs are available that block the secretion of corticosteroids. In addition, drugs have been introduced that block the peripheral action of glucocorticoids by displacing them from the specific receptors. Ectopic corticotropin-producing tumours are treated either by surgery, X-ray therapy, or chemotherapy. Occasionally, if the Cushing's syndrome becomes life-threatening and the usual forms of therapy have been unsuccessful, both adrenal glands may have to be removed. The ensuing adrenal insufficiency is treated in a fashion similar to that in patients with Addison's disease. In those patients in whom the primary cause is a pituitary tumour, bilateral adrenalectomy is sometimes followed by a rapid progression in growth of the tumour along with intense skin pigmentation, a combination known as Nelson's syndrome.

Hypoaldosteronism. Total destruction of the adrenal glands by definition includes hypoaldosteronism as part of the disorder. There exists, however, a disease in which adrenocortical function is intact except for defective synthesis and secretion of aldosterone from the zona glomerulosa.

Isolated aldosterone deficiency results in a low level of sodium in the serum (hyponatremia) along with an elevated level of potassium (hyperkalemia). These biochemical changes produce weakness as well as an increased risk of dangerous abnormalities in heart rhythm, some of which are fatal. Hypoaldosteronism is frequently associated with kidney disease, especially in diabetics, and in these instances the cause stems from deficient production of renin with consequent low levels of angiotensin and a reduced stimulus for the secretion of aldosterone. Rarely, the deficiency lies in an inadequate production of the enzyme needed to synthesize angiotensin (angiotensin-converting enzyme). As a result, plasma renin levels are elevated. In other cases there is an enzymic defect in aldosterone production, which may be hereditary. Treatment requires the administration of fludrocortisone, a powerful synthetic mineralocorticoid. Aldosterone itself is poorly absorbed when taken orally.

Hyperaldosteronism (Conn's syndrome). In 1955, an American internist, Jerome Conn, described a form of high blood pressure associated with hypokalemia and reduced acidity of the blood (alkalosis) in patients who harboured a benign tumour of adrenal glomerulosa cells. These patients were found to have high levels of aldosterone in the circulation, and for most the hypertension and hypokalemia disappeared with the removal of the adrenal adenoma. Aside from the sometimes severe symptoms of high blood pressure, such as headache, patients often note weakness, increased urination, increased thirst, and peculiar skin sensations along with muscle cramping. Abnormalities in heart rhythm also may occur, and if potassium loss is severe there may be impaired glucose tolerance, although diabetes is not common.

Symptoms of Conn's disease

Hyperaldosteronism may occur as a secondary phenomenon in other diseases, particularly those accompanied by increases in extracellular fluid (edema). Examples include heart failure, severe liver disease, and a kidney ailment, nephrosis, characterized by excessive loss of plasma proteins. While the cause of increased aldosterone secretion in these illnesses is not clearly understood, successful treatment of the primary disease leads to a restoration of aldosterone levels to normal.

An American endocrinologist, Frederic Bartter, described individuals who exhibited hyperplasia of the juxtaglomerular apparatus, high serum renin and angiotensin levels with resultant elevations in plasma aldosterone associated with hypokalemic alkalosis. These individuals, however, had a consistently normal blood pressure. The onset is usually in late infancy or childhood, and patients often show evidence of dwarfism and mental retardation. The cause is not well understood, but the hypokalemia and some of the symptoms may be reversed by the use of drugs, such as indomethacin, that inhibit the formation of prostaglandins.

Congenital adrenal hyperplasia. Congenital adrenal hyperplasia is a disorder in which the hereditary absence of a single enzyme has far-reaching consequences. In the most common form of this deficiency, an adrenal enzyme called 21-hydroxylase is absent. As a result, the adrenals cannot synthesize aldosterone and cortisol. The low levels of circulating cortisol reduce inhibition of corticotropin secretion by the pituitary. The resulting high levels of corticotropin lead to excessive secretion of adrenal androgens. When this enzyme deficiency is absolute, the child may die at, or soon after, birth from adrenal insufficiency. When the enzyme deficiency is only partial, the child may survive. Because the excess of adrenal androgens begins in utero, however, children are born with striking signs of masculinization (virilization): newborn genetic females have an enlarged clitoris, often mistaken for a penis, and an enlarged vulva, which resembles a bilobed scrotum. These individuals, known as female pseudohermaphrodites, may reach maturity and live out their lives as short, stocky males. They are, of course, infertile since they have vestigial ovaries rather than testes. A variation on this disorder occurs late in adolescence and is diagnosed in women who appear normal except for the development of excessive hair on the face and extremities (hirsutism). In genetic males, the excessive androgens lead to striking muscle development and an enlarged penis, the "infant Hercules."

21-hydroxylase deficiency

Treatment of the juvenile form of this disorder depends upon the time of diagnosis. If the patient is near the age of puberty, it is generally considered wise to permit the genetic female to maintain the male gender role since it has become deeply embedded. When the diagnosis is

made at birth, however, treatment with replacement doses of cortisol permit a reversal of the entire process. Normal levels of cortisol reduce the excessive secretion of corticotropin, which in turn decreases the secretion of male sex hormones (adrenal androgens) to normal. The patient then develops normally, and the ambiguous genitalia can be corrected surgically. For the late-onset type, treatment with cortisol or one of the synthetic glucocorticoids arrests the process.

Other enzyme deficiencies in adrenal hyperplasia result in still other dramatic variations. The absence of 17α-hydroxylase leads to a stockpiling of steroid precursors, sometimes including a powerful mineralocorticoid called desoxycorticosterone. The result in a child is similar to that seen in primary aldosteronism (hyperaldosteronism) with hypertension and hypokalemic alkalosis. Another genetic defect, 18-hydroxylase deficiency, blocks the formation of aldosterone so that the child shows evidence of mineralocorticoid deficiency, excreting excessive amounts of salt. This is a hereditary form of hyperaldosteronism. These variants are also treated with replacement doses of the deficient hormone.

THE ADRENAL MEDULLA

Anatomy. The adrenal medulla is embedded in the centre of the adrenal cortex. It is quite small, making up only about 10 percent of the total adrenal weight. It is composed of chromaffin cells, so called because the granules within the cells darken after exposure to chromium salts. Chromaffin cells have migrated from the embryonic neural crest and represent specialized neural tissue. Indeed, the adrenal medulla forms an integral part of the sympathetic nervous system, a major subdivision of the autonomic nervous system (see NERVES AND NERVOUS SYSTEMS: *The autonomic nervous system*), and the combined activities have been referred to as the sympathoadrenal system.

Medullary hormones

Included among the medullary hormones, the catecholamines, are dopamine, norepinephrine, and epinephrine, all of which are synthesized in the brain and sympathetic nerve endings. The adrenal medulla differs from most other endocrine glands in that the major stimulus for the release of the catecholamines is by stimulating sympathetic nerve endings to release acetylcholine (ACh), an important neurotransmitter of the peripheral nervous system (nerves and ganglia located outside the central nervous system, or the brain and spinal cord). When stimulated, the medullary cell ejects the chromaffin granules from the cytoplasm into the bloodstream, a process known as exocytosis. Thus, the adrenal medulla is a neurohemal organ.

Catecholamines. The catecholamines are synthesized from the amino acid L-tyrosine. Serial changes in chemical structure are catalyzed by enzymes, leading to the following synthetic sequence: L-tyrosine → L-dopa (dihydroxyphenylalanine) → dopamine → L-norepinephrine (noradrenaline) → L-epinephrine (adrenaline). The close proximity of the adrenal cortex to the adrenal medulla is not accidental. The enzyme that mediates the transformation of L-norepinephrine to L-epinephrine is formed only in the presence of high local concentrations of glucocorticoids from the adjacent cortex; chromaffin cells in tissues outside the adrenal cortex are incapable of synthesizing epinephrine.

Synthesis of catecholamines

L-dopa is well known for its role in the treatment of parkinsonism, but its biological importance lies in the fact that it is a precursor of dopamine, a neurotransmitter widely distributed in the central nervous system, including the basal ganglia of the brain (groups of nuclei within the cerebral hemispheres that collectively control muscle tone, inhibit movement, and control tremor). It is a deficiency of dopamine in these ganglia that leads to parkinsonism, a deficiency that is at least partially repaired by the administration of L-dopa. Under ordinary circumstances, far more epinephrine than norepinephrine is released from the adrenal medulla; in the catecholamine neurotransmitting function throughout the body, norepinephrine is far more widespread. It is likely that the full complement of hormones secreted by the adrenal medulla is not yet completely known. There is strong evidence to indicate, for example, that enkephalins (neurotransmitters with opiate-like effects) are contained within chromaffin granules and are secreted into the general circulation.

In physiological terms, a major action of the hormones of the adrenal medulla conjoined with the sympathetic nervous system is to initiate a rapid, generalized bodily response described by Walter Cannon as "fight or flight." This response may be triggered by a fall in blood pressure, pain (including burns), or abrupt emotional upheavals. An injection of epinephrine, in fact, closely mimics the symptoms of an anxiety attack (sweating, tremor, greatly increased heart rate). Metabolic changes also stimulate catecholamine secretions as evidenced by the rapid rise in plasma epinephrine levels when an individual becomes hypoglycemic (has a greatly decreased glucose level). Thus, much of what is called a hypoglycemic reaction is the result of a large epinephrine discharge.

Action of catecholamines

The action of catecholamines on the body's organs and tissues is widespread and complex. There actions, however, are rarely isolated; they usually occur in concert with other neural or hormonal responses. Furthermore, tissue responses depend on the fact that there are two major types of adrenergic receptors on the surface of target organs and tissues: alpha-adrenergic and beta-adrenergic receptors, or alpha receptors and beta receptors, respectively (see NERVES AND NERVOUS SYSTEMS: *Biodynamics of the vertebrate nervous system*). Both contain a number of subtypes so that receptor responses to the catecholamines have some degree of specificity and coordination. In general, the alpha-adrenergic receptors constrict blood vessels and the uterus, relax the intestine, and dilate the pupils. Beta receptors stimulate the heart, dilate the bronchi and blood vessels, and relax the uterus and intestines. It should be noted that there also are specific receptors to dopamine.

The effects of catecholamines on the heart result mainly from their association with beta receptors. When catecholamines bind to these receptors in the surface membranes of heart cells, pulse rate and strength of heart muscle contraction are increased so that the amount of blood moved through the heart per minute (cardiac output) increases. This sequence of events increases the body's requirement for oxygen and raises blood pressure, a consequence of greater blood flow through the heart. Drugs, such as propranolol, that block the activation of these beta receptors (beta blockers) are often used for the treatment of high blood pressure and cardiac pain (angina pectoris). Conversely, since activation of beta receptors results in bronchial dilation (expansion of the air passages in the lung) because of the heightened need for oxygen, propranolol is contraindicated in the treatment of asthma; it would worsen the bronchial constriction that already exists in the condition.

Catecholamines also play key roles in the generation of body heat (thermogenesis). Oxygen is consumed in metabolic processes in the body that produce heat. When catecholamines stimulate beta receptors and increase the overall level of oxygen in the body, more oxygen is consumed, and more heat is produced. Catecholamines also increase available body fuels such as glucose and free fatty acids. They stimulate the breakdown of glycogen, which is stored in the liver and muscle, to glucose (glycogenolysis) and the breakdown of triglycerides, the stored form of fat, to free fatty acids (lipolysis). Finally, the catecholamines are regulatory agents in hormone secretion; they serve as neurotransmitters modulating the secretion of releasing hormones in the hypothalamus, and they stimulate the release of glucagon and somatostatin and inhibit the release of insulin from the islets of Langerhans of the pancreas.

Role as neurotransmitters

All catecholamine effects on hormonal secretion are stimulatory and affect the thyroid and parathyroid, the gonads (ovary and testis), and the placenta. There is evidence, however, that stimulation of dopaminergic receptors blocks the secretion of aldosterone from the adrenal cortex. Excessive secretion or ingestion of the thyroid hormones increases the number of beta receptors so that many of the clinical consequences of the hyperthyroid state can be suppressed by using beta blockers.

Adrenomedullary dysfunction. Isolated loss of the medulla of both adrenals does not occur; such destruction is always accompanied by impairment of the function of the

cortex of both adrenals. Any effects that can be attributed to the loss of the medulla are overshadowed by the predominating signs of Addison's disease.

Tumours of the adrenomedullary chromaffin cells, called pheochromocytomas, do occur and may produce striking, largely predictable signs and symptoms that are exaggerations of the physiological actions of the catecholamines. Pheochromocytomas are tumours of the chromaffin cell, usually benign but occasionally malignant. Commonly unilateral, these tumours may be present in both adrenals when they appear in the hereditary form of multiple endocrine neoplasia. Extra-adrenal tumours of these chromaffin cells have been found in multiple locations, extending from the patient's neck to the urinary bladder, most often in a collection of cells known as the organ of Zuckerkandl. While the normal adrenal medulla secretes mostly epinephrine, pheochromocytomas predominately secrete norepinephrine.

High blood pressure is an invariable finding in adrenomedullary hyperfunction. It may be constant, and it may be difficult to distinguish from the common forms of hypertension. In some instances, however, there is a sudden increase in norepinephrine secretion, provoking the sudden explosive onset of its vasopressor actions, such as a severe headache, excessive sweating, palpitation of the heart, ashen pallor, tremor, and anxiety. These attacks may end abruptly and the patient may appear to be normal following the attack. They may last from minutes to hours and may occur at intervals ranging from, for example, once a month to several per day. In persons in whom tumours secrete an appreciable amount of epinephrine, anxiety may be more marked and the patient may lose weight, be feverish, and show evidence of diabetes mellitus.

Treatment

Excess secretion of either norepinephrine or epinephrine by such tumours may be treated therapeutically or in preparation for surgery by using alpha- or beta-receptor antagonists (drugs that compete with epinephrine and norepinephrine for receptor sites on target organs but do not elicit a response once bound; in essence they tie up many of the potential binding sites of these overly secreted catecholamines). Surgical removal of the isolated tumour remains the favoured treatment. When malignant pheochromocytomas have spread to other organs, however, antagonist drugs may be continued indefinitely.

THE OVARY

Anatomy. The ovaries are multipurpose organs. They harbour, nurture, and guide the development of the egg so that when it is extruded from the ovary (ovulation) it has been prepared for its migration down the fallopian tube, its penetration by sperm, and its eventual implantation in the wall of the uterus. Additionally, the ovary is a sophisticated endocrine structure. It secretes hormones essential for the onset of menstruation (menarche) and its cyclical perpetuation. At the same time, the ovary produces profound alterations in body physique that transforms a prepubertal girl into a mature woman.

The mature ovary is a roughly bean-shaped structure weighing about 14 grams. It, like the adrenal gland, consists of an outer cortex and a central medulla with the addition of an inner hilus (depression or pit) that serves as the point of entry and exit of blood vessels and nerves. The ovaries are located in the pelvis, attached to a structure called the broad ligament (see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The human reproductive system*).

Follicle maturation

Immature follicles (primordial follicles) embedded in fibrous tissue (stroma) enlarge as the follicle matures and moves through the cortex toward the outer surface of the ovary. The cells lining the follicle multiply and become layered into a zona granulosa. Along with this change the stromal cells immediately surrounding the follicle arrange themselves concentrically to form a theca (an enclosing sheath). This egg-containing mature structure is known as a Graafian follicle. Both granuloma cells and thecal cells secrete steroid hormones known as estrogens. The follicular fluid bathing the ovum is an extraordinarily complex liquid containing not only high concentrations of estrogens but also other steroids (progestogens and androgens), pituitary hormones (FSH, LH, prolactin, oxytocin, and vasopressins), and numerous enzymes and bioactive proteins.

During the maturation (follicular) phase of the menstrual cycle, follicles continue to enlarge until one (or, rarely, two) follicles rupture at the ovarian surface. The egg is extruded and promptly enters the fallopian tube to begin its journey to the uterus. The supportive role of the follicle does not end with the discharge of the egg. Thecal cells penetrate the emptied follicle and, together with persisting but modified granulosa cells, fill the follicle, now called a corpus luteum, which is the source of serum progesterone during the postovulatory (progestational or luteal) phase of the menstrual cycle. With menstruation, the corpus luteum becomes scarred and contracted (atretic), remaining as a corpus albicans. In the event that the extruded egg is fertilized and pregnancy ensues, the corpus luteum persists and continues to secrete increasing amounts of progesterone during the first trimester. As might be expected, these changes are controlled by secretions from the hypothalamus and the anterior pituitary gland.

Regulation of hormone secretion. Before the onset of puberty the ovaries are quiescent, and the stroma of the cortex and medulla are studded with multiple primordial follicles. Puberty is heralded by subtle but far-reaching changes. Some undefined event stimulates the secretion of luteinizing hormone-releasing hormone (GnRH) from the hypothalamus, and GnRH secretion becomes pulsatile. Animal studies support the notion that puberty is precipitated by a reduction in the secretion of melatonin, a hormone of the pineal gland. There is, however, no evidence that melatonin has a role in the onset of puberty in humans.

The onset of puberty

Secretion of GnRH activates gonadotrophs from the anterior pituitary, resulting in enhanced secretion of both follicle-stimulating hormone (FSH) and luteinizing hormone (LH). The secretion of these hormones, particularly LH, is much enhanced shortly after the onset of sleep; increased nocturnal secretion of LH is the earliest change detectable in the pubertal child. It appears that GnRH secretion is inhibited by neurons that secrete dopamine and is stimulated by noradrenergic neurons (involved with norepinephrine). Endogenous opiates, especially beta-endorphin and dynorphin, also play important roles in regulating the frequency and strength of GnRH secretion.

The increased secretion of estrogens from the ovaries, stimulated by LH secretion coupled with maturing Graafian follicles (resulting from the increased FSH secretion) leads to menarche. Before long, the cyclic activity characteristic of the normal female hypothalamus appears (Figure 14B). Immediately following the cessation of menstruation, the sequence begins with a gradual rise in the blood level of estradiol (the most potent of the estrogens), paralleled by a slow rise in serum LH. An inconspicuous rise in androgens also occurs while progesterone and its precursor, 17-hydroxyprogesterone, remain suppressed. Finally, the rising estradiol level trips off a mid-cycle surge of LH and FSH (an example of a positive feedback mechanism). The abrupt rise in gonadotropins precipitate ovulation, ending the follicular phase. With the formation of the corpus luteum, estrogen levels fall but not back to baseline, while the levels of 17-hydroxyprogesterone and progesterone are much elevated. At the end of the luteal phase all hormonal levels return to baseline, and the withdrawal of the estrogens precipitates the next menstrual period. The normal menstrual cycle is 28 days long although it varies considerably from one woman to another and occasionally in the same woman, with irregularities occurring most frequently shortly after puberty or before the menopause.

The premenstrual fall in levels of estrogen and progestins occurs because of a degeneration and loss of function of the corpus luteum (luteolysis) that results from a faltering of LH pulses from the pituitary. The endometrium, which had become increasingly thickened and vascular, undergoes a constriction of small arteries. Cutting off oxygen and nutrient supplies to the endometrial lining leads to cell death and the subsequent sloughing and bleeding characteristic of menstruation. It should be noted that the basal body temperature, which fluctuates only mildly during the follicular phase, shows a rather abrupt progressive rise

Luteolysis

after ovulation, paralleling the increase in progesterone (Figure 14A). This thermogenic action results from the effect of the elevated progesterone levels on temperature-regulating centres in the base of the brain. (The structural and functional changes that occur in the fallopian tubes [oviducts], lining of the uterus [endometrium], distal opening to the uterus [cervix], and vagina and that accompany the endocrinologic fluctuations are discussed in the article REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The human reproductive system.* The extension and accentuation of these changes, which occur in the event of pregnancy, are also discussed there.)

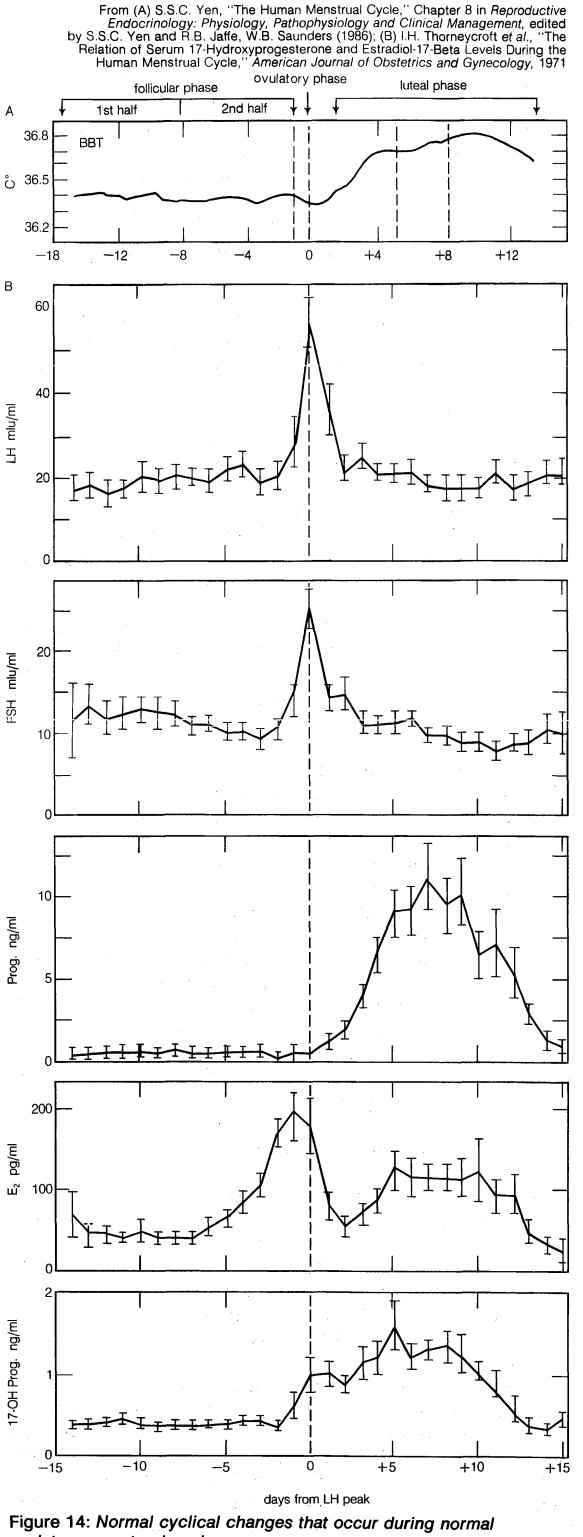


Figure 14: *Normal cyclical changes that occur during normal ovulatory menstrual cycle.*
(A) Changes in basal metabolic temperatures (BBT). (B) Normal levels of luteinizing hormone (LH), follicle-stimulating hormone (FSH), progesterone (Prog.), estradiol (E_2), and 17-OH progesterone (17-OH Prog.).

Hormones. As is the case in the adrenal cortex, the parent sterol from which all ovarian steroid hormones are formed is cholesterol. Both estrogens and progestogens are synthesized from a common precursor, pregnenolone, itself formed from cholesterol. These chemical sequences also include dehydroepiandrosterone, androstenedione, and testosterone, all of which are steroids that are primarily androgens (male sex hormones).

Synthesis of ovarian hormones

Once secreted into the blood, estrogens share with androgens, particularly testosterone, a binding globulin (testosterone-estradiol-binding globulin, TeBG), which transports them to target tissues. At this site, the estrogens easily penetrate the cell surface and are bound to an intracellular binding protein. It is in this form that they are transported to the cell nucleus, where they modulate protein synthesis by influencing the formation of DNA.

Estradiol, the most potent of the three major estrogens, is formed by both granulosa and thecal cells, perhaps acting together. Estrone can be formed from estradiol, but its major precursor is androstenedione. Estriol is formed from both estrone and estradiol and is the weakest of the estrogens. Indeed, in some circumstances estriol appears to have anti-estrogen effects; that is, it may bind to tissue estrogen receptors without setting in train estrogen effects and, while doing so, block the receptor from access to more potent estrogens such as estradiol. It has been suggested that it protects against the development of breast cancer in women. Catechol estrogens are metabolic products that also have anti-estrogen effects.

Formation of progesterones

The progesterones (progestins) are formed by the corpus luteum. Progesterone is also produced by the adrenal cortex, but the rise that occurs in its serum level during the luteal phase stems from ovarian secretion. The hormone 17-hydroxyprogesterone is secreted by thecal cells and accounts for most of the hormone found in the blood; again, the adrenal cortex may secrete a lesser amount at a constant rate.

Among the many nonsteroidal substances secreted into the follicular fluid is a substance called inhibin (folliculostatin), which is secreted by granulosa cells (and by Sertoli cells in the male). The primary action of inhibin is to inhibit the secretion of pituitary FSH. Since the major action of FSH is to stimulate the formation and function of granulosa cells, the relationship between inhibin and pituitary FSH represents a classical negative feedback servomechanism. Relaxin, a polypeptide hormone produced by the corpus luteum, induces a relaxation of the pubic ligaments connecting the two halves of the pelvis, an action that mitigates the discomfort of a woman in labour and eases the passage of the child. Finally, the ovary contains both oxytocin and vasopressin in high concentration, which serve a paracrine function. Oxytocin may assist in the expulsion of the egg from the ovary and may also mediate the process of luteolysis (break up of the corpus luteum). Vasopressin constricts local blood vessels after the egg is extruded.

Functions of ovarian hormones

Ovarian hormones have multiple functions. Pulsatile secretion of LH occurs well before the onset of the first menses (menarche) so that the rate of estrogen secretion also increases progressively. This results in the progressive development of breasts (thelarche) and the appearance of pubic hair and culminates in the menarche. Estrogens, including those contained in oral contraceptives, also exert ongoing generalized effects in the adult female. They mildly impair the body's ability to metabolize glucose, and they tend to increase the level of fats (triglycerides) in the blood. These effects are easily obviated by other endocrine adjustments in the normal woman, but they have an impact when these compensatory mechanisms are impaired. Estrogens increase the serum concentration of a large number of binding proteins that transport other materials; these include binding proteins for cortisol, thyroxine, iron, and copper, as well as those that bind estrogens and testosterone (TeBG). Finally, estrogens tend to increase the concentration of sodium, and therefore the degree of water retention, again particularly in susceptible women.

Diseases and disorders. *Precocious puberty.* In a setting such as the ovary, in which a complex train of interlocking hormonal activities must occur in proper sequence, it is not surprising that there are a number of abnormalities in function. Among them is precocious puberty, defined as the onset of menstruation before the age of nine years. In true precocious puberty, which may occur as early as the age of two years, ovulation takes place, and the child must be protected from the threat of pregnancy. The cause of this disorder is unknown, and affected girls are otherwise normal.

Nonetheless, treatment is important for proper psychological and social development. Synthetic steroids that inhibit the secretion of gonadotropins from the pituitary have been used successfully. They lead to the regression of breast size, the suppression of menstruation, and the prevention of aberrations of height (a pubertal growth spurt that takes place too early, followed ultimately by short stature in the adult). This form of treatment is being superseded by the use of synthetic derivatives of hypothalamic GnRH. These derivatives are long-acting and block pituitary receptors, both by occupying receptor sites and by preventing the pulsatile stimulation by the naturally occurring GnRH.

Pseudoprecocious puberty

Pseudoprecocious puberty occurs from tumours that secrete GnRH or from tumours that secrete the gonadotropins themselves. Although these patients show breast development, the appearance of pubic hair, and menstrual bleeding, they do not ovulate and they are infertile. Treatment is directed against the tumour that is secreting the hormones.

Menstrual disorders. In the adult female who has matured normally, any of a number of menstrual dysfunctions can appear, most of which lead to infertility. Indeed, a common disorder that occurs most often during adolescence or around the menopause is dysfunctional uterine bleeding. Menstrual bleeding occurs at irregular intervals and ovulation is absent. It results from a sluggish LH response to rising estrogen levels in girls and a suboptimal estrogen rise in aging women. Treatment may be with progestins or an estrogen-progestin combination. A second disorder that is being diagnosed with increasing frequency is hyperprolactinemia (see above *The anterior pituitary gland: Prolactin*), but not always associated with a disappearance of menses (amenorrhea) and the spontaneous secretion of breast milk (galactorrhea).

A rare, but intriguing, aberration is false pregnancy (pseudocyesis). A woman, wanting to be pregnant, generates impulses from the cerebral cortex that stimulate the hypothalamus, resulting in an increase in the secretion of LH and prolactin. Changes typical of a first trimester pregnancy then appear. These changes regress promptly when the patient is made aware that she is not pregnant. A mild form of this disorder occurs in patients in whom the corpus luteum persists for longer than the normal 14 days. This persistent corpus luteum syndrome may lead to a prolonged period of amenorrhea.

Anorexia nervosa

A more threatening form of amenorrhea occurs in patients with anorexia nervosa, in which voluntary food restriction is stringent, and in those with bulimia, in which periodic eating binges are followed by self-induced vomiting. The severely affected patient, by mechanisms not well understood, physically regresses to a prepubertal state; amenorrhea and suppression of gonadotropin secretion occur. This regression is associated with a reduction in total fat stores below a critical weight and is almost always reversed by weight gain. These changes may be related to the ability of fat tissue to metabolize estrogens from estrone through the most potent steroid, estradiol, to the least potent, estriol.

The treatment of patients with anorexia nervosa varies with the severity of the malnutrition. Intravenous infusion of nutrients (parenteral alimentation) is used in those whose illness has become life-threatening. In those affected less severely, the treatment is psychological, and the best results are obtained when the patient's family participates along with the patient. A less threatening variant is the amenorrhea of the ballet dancer or marathon runner whose single-minded pursuit of excellence through strenuous physical activity leads to a reduction of fat mass below the critical level. Again, if the individual is compliant, effective treatment consists simply of reducing the amount of strenuous physical activity.

The psychological influence on hypothalamic function is also manifest in a far less severe disorder, psychogenic amenorrhea. Individuals so afflicted respond to psychological stress by reducing gonadotropic secretion; normal menses return when the stress is relieved.

Some young adult women who are in otherwise good health, suffer at the time of menstruation from pelvic cramps (dysmenorrhea), headache, nausea, vomiting, weakness, and dizziness long thought to be psychological in origin. It has become clear that the symptoms are not psychogenic but, rather, are due to an increased secretion of prostaglandins from the uterus. Prostaglandins increase the strength of uterine contractions and cause widespread constriction of blood vessels. The patient's symptoms are completely or partially relieved by taking drugs that inhibit the synthesis of prostaglandins (*e.g.,* ibuprofen, or indomethacin).

Premenstrual syndrome

Another constellation of symptoms that has generated debate is premenstrual syndrome (PMS). Perhaps one in 10 women with normal cycles becomes aware of breast tenderness, weight gain with bloating of the abdomen and swelling of the feet, increased irritability with mental depression, and fatigue. The symptoms appear seven to 10 days before the onset of menstruation. In extreme instances, episodes of violence or other forms of psychotic behaviour follow. Evidence suggests that PMS is due to an increased secretion of endogenous opiates, followed by a rather abrupt withdrawal of these mood-altering hormones. A number of drugs, including GnRH analogues, narcotic antagonists, and clonidine, a drug that stimulates alpha-andrenergic receptors, have been used in treatment.

In addition to the functional (secondary) causes of amenorrhea described above, there are abnormalities in which menstruation is never initiated, the primary amenorrheas. These disorders may reflect serious problems in growth and development (see below *Growth and development: Disorders of growth*). Only a more benign aberration is considered here, that of delayed puberty and adolescence. It is not rare that an otherwise normal girl not have the onset of puberty until after age 13, with delays extending as long as age 18. This delay in adolescence is a benign variant of the normal pubertal process. It is important to reassure the patient and her family and to deal with the psychological problems that this lag in development may produce.

Physical effects of menopause

The menopause. The menopause occurs in women between the ages of 45 and 55 years, with the five-year interval between 45 and 50 being the most frequent time of onset. During the menopause the ovary's eggs and their nurturing Graafian follicles are depleted, and the ovary shrinks and becomes wrinkled; it contains many atretic follicles, and the stroma cells become much increased. Such ovaries cannot synthesize sufficient estrogen to sustain the premenopausal relationships of the hypothalamic-pituitary axis. Thus, as the serum estrogen levels fall, pituitary secretion is less inhibited, and there is a progressive increase in the levels of pituitary gonadotropins. The senescent (aging) ovary, however, is no longer capable of responding to pituitary hormones.

As a result of progressive estrogen deficiency, the uterus and breasts decrease in size, the vagina becomes dry, and sexual intercourse often becomes painful (dyspareunia). In about three-fourths of all females there occurs some degree of increased irritability along with "hot flashes," characterized by a flushing of the skin, profuse sweating, and a feeling of warmth. These episodes are a frequent source of embarrassment during the day and a cause of insomnia at night. While the mechanism is not clearly understood, it has been shown that there is an abrupt rise in pituitary LH secretion simultaneous with a rise in skin temperature. Probably both of these parallel changes are due to simultaneous stimulation of temperature-regulating centres and GnRH production at the base of the brain. There is a minority of women who suffer no menopausal symptoms. It is thought that they are able to convert

enough steroid precursors into estrogens, particularly in fat tissue, to avoid the abrupt fall in estrogen levels that are the rule. An important consequence of the menopause in caucasian women is osteoporosis (see above *The parathyroid glands: Metabolic bone disease*).

Premature ovarian failure

When menses ceases before age 40, the patient is said to have premature ovarian failure. Some of these patients may be afflicted with a genetic disorder. In others, premature failure may be generated by ovarian autoantibodies. Others suffer a loss of ovarian receptors (the resistant ovary syndrome) for unknown reasons.

Administration of estrogens suppresses menopausal symptoms in most women, but estrogen treatment increases the risk of carcinoma of the lining of the uterus and perhaps of the breast as well. Some physicians recommend a combination of estrogens and progestins to simulate a normal menstrual cycle. Some investigators have recommended that all menopausal women be given hormonal replacement indefinitely, but others are concerned that the prophylactic administration of these hormonal agents to the entire postmenopausal population may be costly in terms of benefits as compared to risks. They believe that the administration of hormones should be reserved only for those who have distressing symptoms or who are likely to develop osteoporosis.

Functional androgen excess. Androgens are integrated into the normal endocrinologic pattern of functioning in the adult woman. The two major androgens secreted by the ovaries are androstenedione and testosterone; dehydroepiandrosterone (DHEA) and DHEA-sulfate (DHEAS) are contributed by the adrenal cortex. Other tissues, including skin, fat, muscle, and brain, are capable of converting precursor steroids locally to active hormones, thus permitting the accumulation of high concentrations of steroids in key local areas without having a generalized effect throughout the body.

Local activation of steroid hormones

In view of the numerous sites of androgen production, it is not surprising that there are multiple causes for syndromes of androgen excess. Some, such as Cushing's syndrome, congenital adrenal hyperplasia, and androgen-producing adrenal tumours, have been discussed previously (see above *The adrenal cortex: Congenital adrenal hyperplasia*), where the important distinction between hirsutism and virilization is made. Tumours (including cancers) of granulosa and thecal cells of the ovary usually overproduce both estrogens and androgens and may result in all of the features of androgen excess.

While it is important to rule out these serious illnesses, they are relatively rare. Hirsutism is common, however, occurring most often as part of the syndrome of polycystic ovaries (PCO). Since any of a number of androgens may be secreted in excess in this syndrome, it probably has more than one cause; and the androgens arise either from the adrenals or the ovaries. The ovaries become enlarged with a thickened capsule and contain many atretic follicles. Typically, the patient has hirsutism, amenorrhea, infertility, acne, and obesity. The obesity leads to an excess of estrogens from conversion of androgens in peripheral tissues, resulting in impaired secretion of gonadotropins and a consequent suppression of ovulation, with or without amenorrhea.

Treatment involves suppressing excess adrenal androgen production by using synthetic glucocorticoids such as prednisone, by suppressing the actions of the ovary using oral contraceptives, or by blocking the androgenic receptors in tissues with antagonistic drugs such as spironolactone. Effective treatment restores normal menstrual activity and fertility, and although it does not reverse the hirsutism in most instances, further progression is arrested.

THE TESTIS

Anatomy. The testes, or testicles, are the male gonads. They contain germ cells that differentiate into mature spermatozoa, supporting cells called Sertoli cells, and testosterone-producing cells called the Leydig cells. The germ cells migrate to the fetal testes from the embryonic yolk sac. The Sertoli cells are analogous to the granulosa cells in the ovary, and the Leydig (interstitial) cells are analogous to the stromal cells of the ovary.

The embryonic differentiation of the primitive, indifferent gonad into either the testes or ovaries forms a fascinating chapter in fetal development (see below *Growth and development*). Testosterone and its potent derivative, dihydrotestosterone, play key roles in the formation of male genitalia in the fetus in the first trimester of pregnancy. During the first four weeks after birth, they sensitize the genitalia to respond appropriately to androgens when puberty begins. The testes are formed in the abdominal cavity and descend into the scrotum during the seventh month of pregnancy. Stimulation of testicular descent is provided by androgens, along with a protein hormone called Müllerian-inhibiting substance. It is not uncommon in normal males for the testes to be incompletely descended and easily retracted into the abdomen, but this condition usually corrects itself by the age of three months.

Masculinization of the fetus

The adult testis consists largely of a series of tubules with a central cavity. Sperm cells are continuously maturing as they move from the outer edge of the tubule into the central lumen; the most primitive forms, called spermatogonia, differentiate first into spermatocytes and then spermatids. They eventually mature into spermatozoa and are released into the lumen. Spermatozoa travel through the tubular network to be stored in seminal vesicles and, finally, to be ejaculated with the semen. Interspersed among the seminiferous tubules are Sertoli cells, and in the area between tubules (interstitium) are located the hormone-secreting Leydig cells.

Regulation of hormone secretion. Androgen levels in the circulation are regulated by the classical hypothalamic-pituitary-target gland axis (Figure 3). The secretion of pituitary LH (sometimes referred to in the male as interstitial cell stimulating hormone, or ICSH) is secreted following stimulation by gonadotropin-releasing hormone (GnRH) from the hypothalamus. Luteinizing hormone stimulates the Leydig cells to secrete testosterone. When testosterone levels rise above normal, GnRH and LH secretion are inhibited. In the normal course of events, therefore, testosterone levels remain within normal bounds.

The hypothalamic component of this axis comes into play when it is appropriate to override the usual constraints. It has been shown in primates, for example, that serum testosterone levels rise when males are placed in proximity to receptive females, but the level falls when these same males are caged with unreceptive, hostile males to whom they are strangers. It is thought by some that the reduction in serum testosterone levels in such an alien environment is accompanied by a decrease in aggressive behaviour, which, literally, may have survival value. A relation between androgen levels and aggressive behaviour in humans remains uncertain; complex social and interpersonal factors make interpretation difficult.

Transport of testosterone

Like other steroid hormones, testosterone is transported in the plasma bound to a testosterone-binding globulin (TeBG) and to albumin. Only about 2 percent of testosterone is transported unbound in the plasma. Free testosterone is in equilibrium with that which is bound so that when the free steroid enters the cell some bound testosterone is freed simultaneously.

Hormones. Testosterone serves as a circulating prohormone for an important steroidal metabolite, dihydrotestosterone, that performs most of the androgenic functions in the body. Testosterone may also be converted into the potent estrogen estradiol in tissues, particularly adipose tissue. Furthermore, testosterone is interconvertible with androstenedione, which, again in adipose tissue, may be converted to the estrogen estrone.

Testosterone has two major actions: it serves as the feedback inhibitor of GnRH secretion from the hypothalamus and LH secretion from the pituitary, and it directs the development of embryonic Wolffian ducts into the formation of seminiferous tubules. Dihydrotestosterone is responsible for ongoing sperm maturation (spermatogenesis), for the virilization of the embryonic genitalia, and for sexual maturation at puberty. In addition, androgens are powerful anabolic hormones; that is, they enhance the growth of body tissues, particularly muscle.

Normal spermatogenesis requires the secretion of LH and FSH. Luteinizing hormone stimulates testosterone se-

cretion from Leydig cells in the stroma of the testis; the testosterone is converted to dihydrotestosterone, and it must be present locally in high concentration for normal generation of sperm to proceed. Follicle-stimulating hormone acts directly on the seminiferous tubules to stimulate the normal maturation of sperm. Finally, as indicated previously, androgens stimulate Sertoli cells to secrete inhibin. When released into the blood, inhibin dampens pituitary FSH secretion, an additional component of the feedback control mechanism.

Diseases and disorders. *Precocious and delayed puberty.* Male children also can undergo true precocious puberty or the various forms of pseudoprecocious puberty. In addition, there is a poorly understood form of sexual precocity that is a familial (autosomal dominant) disorder in which precocious puberty appears in males in the absence of any increased pituitary gonadotropin secretion and in the absence of any hormone-secreting tumour. The reasons for this inherent premature overactivity of the testes are unknown. It has been suggested that it be called familial testotoxicosis.

Familial precocious puberty

The young child develops pubic hair, a pigmented scrotum, an enlarged penis, and increased muscle development. Since in the affected male child, unlike the female with true precocious puberty as described above, the hypothalamus and pituitary are not activated, treatment with GnRH analogues is not effective. Instead, these patients are treated with an inhibitor of testosterone synthesis called ketoconazole. Severe disease of the testes can prevent completely, or block partially, the onset of puberty.

Hypogonadism. In addition to the functional changes akin to those previously described in females and leading to delayed puberty and adolescence, there are organic diseases that lead to permanent gonadal deficiency. Aside from the various forms of disease and injury, including surgery, that lead to panhypopituitarism, the most common form of gonadotropin deficiency is due to a defect in the hypothalamus that results in an inability to synthesize and secrete GnRH. Persons with this defect (Kallmann's syndrome) may be born with other malformations, including a much undersized penis (microphallus), and there is often an associated loss of smell (anosmia). They do not undergo puberty unless treated. That the defect lies in the hypothalamus is shown by the fact that when these individuals are treated with GnRH, serum gonadotropin levels increase. Conventional treatment of this disorder has been with injections of testosterone, but nasal insufflation of GnRH has been successful.

There are a number of other causes for gonadotropin deficiency. Various tumours in the area of the hypothalamus and pituitary as well as a number of rare disorders, such as the Prader-Labhart-Willi syndrome, may produce hypogonadotropic hypogonadism, in the latter instance due to a chromosomal defect.

Factors causing hypogonadism

In the adult male hypogonadism can occur as a consequence of hypothalamic or pituitary deficiencies (see above *The hypothalamus: Gonadotropin-releasing hormone*). In addition, gonadotropin secretion may be suppressed when hyperprolactinemia occurs (see above *The anterior pituitary: Prolactin*). The testes are susceptible to acquired diseases as well, the most common being mumps orchitis, usually affecting only one testis (unilateral), but when both testes are infected full-blown hypogonadism and infertility may result. Physical trauma, X-ray therapy, and a number of drugs, including commonly used chemotherapeutic agents for the treatment of cancer, can temporarily or permanently impair testosterone synthesis. Alcoholics who sustain severe liver damage are often estrogenized, but alcoholism is associated with a direct inhibition of testosterone synthesis as well. Finally, gonadal failure is commonly associated with a number of chronic illnesses, particularly kidney disease and sickle-cell anemia.

Symptoms of testosterone deficiency

The symptoms of testosterone deficiency in the adult male include the cessation of hair loss. (Normal testosterone levels are necessary for the usual pattern baldness, which occurs frequently in mature men; the converse is not true, however, and most men who retain a full head of hair also maintain normal circulating androgen levels.) The skin of the hypogonadal male is smooth, with a rather fine wrinkling, particularly in front of the ears and around the mouth. Hair in the pubic area and in the beard becomes sparse. There may be some breast enlargement (gynecomastia), and the hips may become broad and assume a female configuration (gynecoid habitus). The testes become smaller than normal, and they may be insensitive to pain. Affected individuals complain of weakness, usually lose interest in sexual activity, and are unable to achieve an erection or to ejaculate.

Treatment of androgen deficiency caused either by a hypothalamic-pituitary axis defect or by a gonadal defect includes the administration of testosterone, usually by intramuscular injection. Many of the symptoms are entirely reversible. Testicular size, however, may decrease even further because of the inhibition of FSH secretion resulting from the administration of testosterone.

Testicular tumours

Hypergonadism. Testicular tumours occur both in children and in adults. Tumours may appear in both testes, and the risk is much increased in those who have undescended testes (cryptorchidism). By far the most common tumours are those of germ cells (seminomas). Previously almost uniformly fatal, testicular tumours are now treated through surgery and chemotherapy, and the survival rates of patients harbouring these tumours have risen dramatically. Tumours of the Leydig cells are quite rare and are almost always benign. Nonetheless, they secrete large quantities of testosterone, precipitating a pseudopuberty in the prepubertal boys and hypergonadism in adult males. The hypergonadal male may lose head hair while showing increased abnormal hairiness in other areas of the body. Acne often reappears, and there may be muscular enlargement due to overgrowth of cells. Because large amounts of the excess testosterone are metabolized to estrogens, patients may develop enlarged breasts.

GROWTH AND DEVELOPMENT

The processes of growth and development are usually accepted as facts of everyday life; however, when one considers the powerful forces at work and the many harmoniously intermingled regulators that harness them, the emergence of a mature adult human being is a source of wonder. The carefully monitored conversion of a crude mixture of nutrients, often ill-balanced, into growing body tissues is integral to the purview of the endocrine system, although the nervous and immune systems play important roles as well.

Patterns of growth

From the 10th to the 20th week of pregnancy, the fetus grows at a rate of 52 inches (132 centimetres) per year. This phenomenal growth rate tapers rapidly as birth approaches. Weight at birth is an important marker. Low birth weight is not surprising in infants coming from families whose histories include low birth weight, but it may also be an indication of premature birth or of poor intrauterine nourishment from a mother living in poverty or with poor hygiene. Growth during infancy remains rapid and then progresses at a slower but steady rate until the onset of puberty, when there is a striking acceleration. The pubertal growth spurt lasts about two years, and it is accompanied by the appearance of secondary sexual characteristics. With puberty, there ensues an increase in nocturnal secretion of growth hormone.

Endocrine influences. Accurate estimates of bone age are made by examining radiographs (a film record of a structure using X rays) of the hands and wrists of large numbers of normal children. In children with endocrine disorders, bone age may not correlate closely with chronological age; bone age is retarded in growth hormone-deficient children and increases in children with growth hormone-producing tumours. Hyperthyroidism, even when it occurs in the developing embryo, is associated with an advanced bone age, while the opposite is true with thyroid deficiency. Children with Cushing's syndrome not only have osteoporosis but retarded bone age as well. An excess both of androgens and of estrogens is associated with a relatively advanced bone age, while a partial androgen deficiency leads to an increase in prepubertal growth of the extremities, resulting in adults with long arms and long legs attached to a short trunk (eunuchoid habitus).

Insulin is a potent growth hormone, and childhood di-

abetics are notoriously small for their ages. Indeed, like hypothyroid children, some never advance to the pubertal state unless proper insulin replacement therapy is provided.

Growth factors. When investigators began studying the effects of biologic materials on cells and tissues developed for laboratory research outside of the body, they discovered a group of peptide hormones that were distinct from any previously known hormones and were active in stimulating the growth in size and number of these cells living outside the body. This group of peptides include somatomedins, epidermal growth factor, platelet-derived growth factor, nerve growth factor, erythropoietin, lymphokines, thymosin, and transforming growth factors, all of which are discussed below.

Somatomedins. The most intensively studied of these peptide growth factors are the somatomedins, also known as insulin-like growth factors. Of these, somatomedin-C (SmC), also called insulin-like growth factor I (IGF I), along with the related insulin-like growth factor II (IGF II) have emerged as the most important biologically. These two somatomedins are distinguishable in terms of specific actions on tissues and, more precisely, different specific tissue receptors. Somatomedin-C is a peptide with an amino acid structure strongly reminiscent of the prohormone of insulin, proinsulin. It is not surprising, therefore, that both somatomedins have effects that mimic those of insulin when incubated with adipose tissue in the laboratory, but they are far less potent than insulin.

Somatomedin-C

The major action of the somatomedins is on cell growth. Indeed, many of the effects of growth hormone are mediated by way of the somatomedins. Growth hormone stimulates many tissues, particularly those of the liver, to synthesize and secrete the somatomedins. The somatomedins, in turn, stimulate both hypertrophy and hyperplasia of most tissues, including bone. In normal children, blood levels of SmC rise progressively through puberty to adolescence. Abnormally low levels of SmC can be found in individuals with growth hormone deficiency, while abnormally high levels of SmC are found in patients with acromegaly. It is likely that the major actions of the somatomedins occur at the site of their formation, where local concentrations are quite high and cell growth can be stimulated without having the somatomedins pass through the general circulation; in effect, the somatomedins and other growth factors may exert their major actions by way of paracrine and autocrine effects.

Epidermal growth factor. Epidermal growth factor (EGF), a peptide containing 53 amino acids, has been found to be identical to urogastrone, a peptide isolated from the urine of pregnant women, which blocks the secretion of gastric juices. An unlikely, but nonetheless major, site of EGF formation is the salivary gland. Epidermal growth factor stimulates many epithelial tissues to proliferate, and it has been postulated that it plays a major role in the rapid proliferation of these tissues in the fetus. In the adult, EGF formation is dependent upon and stimulated by the presence of androgens. The full clinical implications of EGF are uncertain. A study in mice revealed that when the submandibular salivary glands were removed before an adult female became pregnant, subsequent litters had a high mortality rate within the first four weeks after birth and that maternal milk production was greatly decreased.

Platelet-derived growth factor. Platelet-derived growth factor (PDGF) is a polypeptide contained in blood platelets and released during the process of blood clotting. It stimulates the proliferation of fibroblasts (cells essential for healing) at the site of a wound. It may also play a key role in the pathological process called hardening of the arteries (atherosclerosis). Platelets are attracted to and aggregate around collections of fat in the walls of blood vessels (plaques), and the release of PDGF at these sites stimulates the proliferation of cells in the vessel walls, causing them to narrow. The aggregation of platelets has been inhibited by drugs such as aspirin (see below *Prostaglandins*) in the blood vessels of laboratory animals, thus preventing the development of atherosclerosis.

Blood clotting

Nerve growth factor. Nerve growth factor (NGF) plays an important physiological role in fetal life. It stimulates the growth of nerve cells that form the sympathetic nervous system and may play a similar supporting role in normal adults. It has been incriminated in the genesis of two unusual but disabling disorders. The first, called familial dysautonomia, is a serious affliction of the sympathetic nervous system manifested by an inability to sustain blood pressure in the erect posture, along with other defects (see above *The adrenal medulla*). Persons with the disease are unable to synthesize NGF normally. Another rare disorder, intestinal ganglioneuromatosis, which is part of a hereditary endocrine disease, multiple endocrine neoplasia type II (see below *Ectopic hormone and polyglandular disorders*), is characterized by an impressive overgrowth of nerve cells in the intestinal wall thought to be the result of hypersecretion of NGF in local supporting cells.

Erythropoietin. A number of growth factors specific for bone marrow cells have been identified. Chief among these is erythropoietin, which stimulates the bone marrow to increase the production of red blood cells (erythrocytes). Erythropoietin is a rather specialized protein, a sialoprotein, containing 70 percent protein, which is synthesized in the kidney of the adult and is released into the general circulation. In this respect it is a more orthodox hormone since its mode of action is endocrine rather than paracrine or autocrine.

The secretion of erythropoietin is stimulated both by androgens and by growth hormone, but the primary stimulus for erythropoietin secretion is a lack of oxygen in the tissues. The amount of circulating erythropoietin increases greatly in individuals living at high altitude, patients with disease of the heart and lungs, and in those with erythropoietin-producing tumours of the kidney. Erythropoietin deficiency occurs in patients with severe kidney disease and treatment with erythropoietin has been found to alleviate the anemia associated with renal insufficiency.

Mode of action

There are analogous hormonal proteins that are growth factors for two types of white blood cells, both granulocytes and monocytes. These proteins have been referred to as colony-stimulating factors and macrophage growth factors, respectively.

Lymphokines. Lymphocyte production is regulated by growth factors known as lymphokines. Among the lymphokines are a group known as the interleukins. Interleukin-1 stimulates the growth of monocytes and also stimulates the production of interleukin-2, which in turn stimulates the proliferation of T cells. Interleukin-3 acts on lymphocytes at an earlier stage of differentiation (see IMMUNITY).

Thymosin and thymopoietin. The thymus gland has important functions in the immune system, but it also produces a growth factor called thymopoietin, a single-chain peptide consisting of 49 amino acids that stimulates the differentiation of primitive thymus lymphoid cells (prothymocytes) into mature T cells. Thymosin, a 28-amino-acid peptide, stimulates the growth and differentiation of thymocytes and has been reported to be helpful in the treatment of persons with inherited immunodeficiency disease (see IMMUNITY).

The thymus gland

Transforming growth factors. An important area of research has been the exploration of roles played by certain growth factors in transforming normal cells into malignant cells. Unlike normal cells under similar conditions, transformed cells grown in the laboratory in cell cultures multiply at a rapid rate even in the absence of a growth-supporting serum, and, unlike normal cells, which have a limited capacity to replicate in the laboratory, transformed cells become immortal in that they can survive in cell culture indefinitely. Presumably, the transformation permits the cells to synthesize, and be stimulated by, their own growth factors in an autocrine fashion. Transformation can take place in laboratory cultures of cells that have been infected with any of a variety of viruses. While transforming growth factors (TGF) are present in malignant cell cultures, they are present in lower concentrations in normal cells; the synthesis of TGF may be directed by specialized genes, called oncogenes.

Disorders of growth. *Sexual differentiation.* The embryological and anatomic aspects of the gonads and genitalia are detailed in the article REPRODUCTION AND

REPRODUCTIVE SYSTEMS: *The human reproductive system;* and descriptions of chromosomes and the genes they bear is described in GENETICS AND HEREDITY, THE PRINCIPLES OF: *Human genetics,* so that only a brief review is presented here. In humans, each egg contains 23 chromosomes, of which 22 are autosomes and one is a female sex chromosome (the X chromosome). Each sperm also contains 23 chromosomes: 22 autosomes and either one female sex chromosome or one male sex chromosome (the Y chromosome). An egg that has been fertilized by the penetration of a sperm has a full complement of 46 chromosomes, of which two are sex chromosomes. The genetic sex of the individual, therefore, is determined at the time of fertilization; fertilized eggs containing an XY sex chromosome complement are ordained to be males, while those containing an XX array are destined to develop as females.

he sex
ıromo-
ɔmes

Regardless of this preordination, however, all developing embryos become feminized unless masculinizing influences come into play at key times during gestation. A testis-organizing factor assists the Y chromosome in initiating male sexual differentiation by directing the embryonic gonads, which initially are sexually undifferentiated (indeterminate), to develop as fetal testes. The X chromosome also participates in the differentiating process because two X chromosomes are necessary for the development of normal ovaries. In every embryonic life the fetus contains structures capable of developing into either male or female genitalia.

During the third fetal month, the fetal testis of the XY embryo secretes testosterone, an event that has striking consequences. The ducts that would have otherwise developed into oviducts (fallopian tubes) atrophy, while a separate set of ducts (Wolffian ducts) are stimulated to develop eventually into seminiferous tubules along with the ducts (vas deferens) connecting them to the urethra of the penis. If the fetal gonad does not secrete testosterone at the proper time, the genitalia develop in the female direction regardless of whether testes or ovaries are present. In the normal female fetus, no androgenic effects occur; the ovaries develop along with the Müllerian ducts while the Wolffian duct system deteriorates. Sexual differentiation is completed at puberty and a normal adult male or female develops.

In the adult there is a marker for the genetically normal female cell. The nucleus of such a cell contains a darkly staining mass at its edge. This mass, called the X chromatin or Barr body, is an inactive X chromosome. During normal cell activity the DNA of only one X chromosome participates; the other persists only as an inactive Barr body. Such chromatin masses are not found in normal genetic males because they have only one active X chromosome for the cell nucleus. Not the same X chromosome becomes inactive in every cell of a normal female, so that some cells express the X chromosomal activity of the father while others express that of the mother. In effect, every normal female is what is called a mosaic, an individual whose active chromosomal components vary from one cell to another. This state of affairs is known as the Lyon hypothesis.

ɜarr body

It should be mentioned that sexual differentiation occurs in the hypothalamus as well. During the newborn period, exposure to androgen leads to a pulsatile but otherwise unvarying secretion of hypothalamic gonadotropin-releasing hormone throughout adult life. In contrast, the lack of a neonatal androgen influence leads in the mature female to the characteristic monthly cycles of GnRH secretion, reflected in normal menstrual cycles.

In such a complex system there are many opportunities for some form of aberrant development. The causes of these disorders, while not fully understood, have been greatly elucidated by rapid advances in chromosomal analysis, the identification of isolated genetic defects in steroid hormone synthesis, and an expanded understanding of abnormalities in steroid hormone receptors. When techniques became available for microscopic examination of the full complement of individual chromosomes, soon followed by sophisticated fluorescent staining techniques, a good deal of confusion surrounding the clinical distinctions among abnormalities of sexual differentiation were resolved.

Klinefelter's syndrome. Klinefelter's syndrome (47,-XXY seminiferous tubule dysgenesis) is the most frequent chromosomal disorder (occurring in one in 1,000 males). Symptoms and features were first described in 1942 by the American physician Harry F. Klinefelter, a student of Fuller Albright. It later became known that affected individuals had an extra X chromosome in each cell so that the sex chromosome content was XXY and the total number of chromosomes in each cell was 47 rather than 46. Viewed under the microscope, the cells of these individuals contain Barr bodies because, as in normal females, one of the two X chromosomes is inactive.

Symptoms of Klinefelter's syndrome

These patients have the outward appearance of males with firm, small testes. They cannot generate sperm, and they often have enlarged breasts and buttocks and inordinately long legs. Testosterone production is deficient and there is a compensatory increase in the pituitary gonadotropin secretion. While normal in intelligence, some of these persons have difficulties in making social adjustments. Klinefelter's syndrome occurs more often in the children of mothers over the age of 35 years.

The mosaic form of Klinefelter's syndrome (46,XY/47,-XXY) is the second most common type of chromosomal disorder in males. Such persons generally have fewer symptoms than do patients with the complete syndrome. Far rarer variants occur, including 48,XXYY; 48,XXXY; 49,-XXXYY; and 49,XXXXY. These patients suffer from a variety of additional abnormalities and, unlike those with classical Klinefelter's syndrome, they are always mentally retarded. Another variant is the XX male syndrome. Such persons show changes typical of Klinefelter's syndrome. Apparently they have Y chromosome material transferred to one of the autosomes. Treatment with androgens serves to reduce the gynecomastia and evidence of male hypogonadism while increasing the strength and libido of all variants of Klinefelter's syndrome.

Turner's syndrome. Turner's syndrome (gonadal dysgenesis) occurs as a result of a deletion of a sex chromosome so that in the typical patient there is a 45,X chromosomal complement. In genetic terms, these individuals are neither male nor female since the second, sex-determining chromosome is absent. Without a Y chromosome to direct fetal gonads to the male configuration, they develop as females with no Barr bodies demonstrable in cell nuclei. Clinically, they tend to resemble one another—with a small chin and prominent folds of skin at the inner corners of the eyes (epicanthal folds), low-set ears, a short neck with redundant skin (webbed neck), a shieldlike or square chest, and short stature. Both the internal and external genitalia are infantile, and the gonads are present only as "streaks" of connective tissue.

If untreated these patients fail to develop secondary sex characteristics, and they are susceptible to a number of threatening congenital abnormalities of the heart and large blood vessels. Turner's syndrome, in genetic terms, is extremely common since one-tenth of all spontaneously aborted fetuses have a 45,X constitution; only 3 percent of afflicted fetuses survive to term.

Treatment of Turner's syndrome

It is possible to use growth hormone to increase ultimate height of patients with Turner's syndrome, but it is more important to treat them with estrogens at the time of puberty. This leads to the appearance of secondary sexual characteristics along with monthly vaginal bleeding simulating a menstrual cycle. Aside from the psychosocial benefits, estrogen treatment prevents the emergence of the severe osteoporosis found in untreated patients.

As with Klinefelter's syndrome, Turner's syndrome has multiple variants in its chromosomal constitution, which include mosaics and chromosomal translocations (in which a portion of one chromosome is transferred and attached to one of the arms of another chromosome). A frequent variant is the 45,X/46,XY mosaic, in which an individual may be reared as either a male or a female because the genitalia are "ambiguous," it being difficult to determine whether the phallus is an enlarged clitoris or a small penis. These patients also have streak gonads with an increased risk that they will undergo a malignant change.

Rarely, patients with 46,XY or 46,XX chromosome complements are found to have streak gonads, and they never develop secondary sexual characteristics, although they are spared the skeletal changes associated with Turner's syndrome.

True hermaphrodites

Hermaphroditism. Hermaphroditism is, in strict medical terms, quite rare. A true hermaphrodite is an individual who harbours ovarian and testicular tissue, both clearly defined when examined under a microscope. Separate ovarian and testicular tissue may be present, or the two tissues may be combined in an ovotestis. Most often, but not always, the chromosome composition is 46,XX, and in every such individual there also exists evidence of Y chromosomal material on one of the nonsex chromosomes (autosomes). These individuals usually have ambiguous external genitalia with a sizable phallus so that, generally, they are reared as males. They develop breasts during puberty and menstruate. In some instances even pregnancy and childbirth have occurred. Spermatogenesis is rare.

Treatment depends upon the age at which the diagnosis is made. If it is decided that a male identity is deeply embedded and therefore a male role is preferable, all female tissues, including the oviducts and ovaries, are removed. In those persons to be reared as females, the male sexual tissues are removed. In older patients, the accepted gender should be reinforced by the appropriate surgical procedures and hormonal therapy.

There exist rare sex chromosome abnormalities that are not associated with any gonadal defects. These include 47,XXX; 48,XXXX; and even 49,XXXXX. People with such abnormalities are usually mentally retarded, and the diagnosis is often made by the finding of multiple Barr bodies in the nuclei of cells of patients confined in mental hospitals. Males with a 47,XYY complement were long thought to be predestined to become tall men with severe acne who commit violent crimes (XYY syndrome). Later studies have documented that these predictions were greatly exaggerated. Although there appears to be a somewhat increased risk of aberrant behaviour, the majority of such men behave in an entirely normal fashion.

Hormonal causes

Female pseudohermaphroditism. Genetic females (46,XX) who often are assigned the male gender have in the past been produced by hormones used to sustain pregnancy. If, in the first trimester of pregnancy, a mother is administered androgens, progestogens, anabolic steroids such as Danazol, or even the synthetic estrogen stilbestrol, her female child may be masculinized during fetal development. Androgen-producing tumours of either adrenal or ovarian origin may also lead to masculinization of the female fetus. (For discussion of female pseudohermaphroditism due to an enzyme defect in the adrenal cortex, see above *The adrenal cortex.*)

Male pseudohermaphroditism. Male pseudohermaphrodites are genetic males (45,XY) who develop female configurations and identities. The gonads are testes, but the genital ducts and external genitalia are female. Secondary sex characteristics may never appear in some patients, while others may achieve a fully feminized physique. Male pseudohermaphroditism is rare and almost always results from genetic defects, usually autosomal recessive in type. Although a number of specific defects lead to feminization of a genetic male, they all share, by one mechanism or another, a loss of androgenic effects on body tissues. In a few rare instances Leydig cells are absent or greatly reduced in number, presumably because the receptors for LH are defective; without Leydig cells little testosterone is produced. In other patients there are enzymic deficiencies analogous to what occurs in female pseudohermaphrodites (see above *The adrenal cortex*), but in this instance resulting in fetal androgen deficiency.

Complete testicular feminization

In some patients, tissue receptors for androgens are absent or reduced, forming a spectrum of syndromes of partial to complete resistance to androgens. Perhaps the most striking example is complete testicular feminization. Affected patients are born with female genitalia and a vagina that ends blindly. They have well-defined testes located either in the labia or within the abdomen; nevertheless, they grow into well-proportioned, attractive females with normal breasts and scant or absent axillary and pubic hair. They have a strong female orientation, but they do not menstruate. The hormonal aberrations in these patients are dramatic and predictable. With a loss of hypothalamic and pituitary androgen receptors there is no inhibin of gonadotropin secretion, and plasma LH levels remain elevated and lead to enhanced stimulation of the Leydig cells. In consequence, serum testosterone levels are much elevated, and Leydig cells are greatly increased in number. The FSH levels are usually normal, probably due to increased inhibin production by Sertoli cells. The peripheral conversion in tissues of the increased amounts of testosterone to estrogens leads to an increase in estrogen levels above normal values for males.

In another extraordinary variant, the lesion lies not in the loss of androgen receptors but rather in a loss of the 5 α-reductase, an enzyme necessary for the conversion of testosterone to the more potent hormone dihydrotestosterone. In this syndrome, because of a lack of testosterone directing fetal development toward a normal male configuration, genetic males are born with what appears to be female genitalia with an enlarged clitoris. These persons are often raised as females, but at puberty an increase in testosterone secretion leads to clear-cut masculinization without enlarged breasts. There then ensues a transition from a prepubertal female to an adult male. This change in gender identity takes place apparently without undue emotional turmoil.

Vanishing testes syndrome

In some fetuses there occurs, for unknown reasons, a regression and disappearance of the testes of genetic males, the "vanishing testes syndrome." When this occurs early in pregnancy and before androgen-induced differentiation toward male genitalia, the child is born with female genitalia. If the testes disappear during the crucial period between eight and 10 weeks of gestation, the child is born with ambiguous genitalia, whereas if the disappearance occurs after this key period, the individual is a male, but without any testes (anorchia).

Treatment of such persons must be highly individualized. In most instances, the gender identity has been firmly implanted by the age of 18 months, and sexual changes are attempted only after careful consideration. Intra-abdominal testes should be removed because of an increased risk of tumour formation. The patient can be treated at the appropriate time with sex hormones.

Homosexuality and transsexualism. The genesis of homosexual and transsexual behaviour is complex and poorly understood. Undoubtedly, environmental and psychosocial influences play important roles, but only the rather meagre knowledge of endocrine influences is discussed here. While early studies suggested a number of abnormalities in homosexual males, including low serum testosterone levels and abnormal ratios of several steroid hormones, later, more stringent investigations generally have not confirmed these differences. It is clear that treatment of male homosexuals with androgens may increase the sexual drive but only in the direction that had been accepted previously.

Recent theories

There is more recent evidence from studies in animals and from inferential studies in humans that severe emotional stress in mothers early in pregnancy may lead to homosexuality in their male offspring. In some studies, elevated serum testosterone levels were found in female homosexuals, while in others no differences were found. Generally, endocrine function has been found to be normal in transsexual men; however, some studies have indicated that these individuals have mildly elevated serum LH levels along with a hyperresponsiveness to the stimulation of LH by GnRH.

It may well be that these confusing conclusions result from the study of heterogeneous populations. Some homosexual behaviour may be predetermined by aberrant hormonal influences during pregnancy while others may be a response to environmental influences. If so, divergent, indeterminate results of hormonal studies would not be surprising.

THE PINEAL GLAND

The pineal gland, the most enigmatic of endocrine organs, has long been of interest to anatomists. Several

millennia ago it was thought to be a valve that controlled the flow of memories into consciousness. René Descartes, the 17th-century French philosopher-mathematician, concluded that the pineal was the seat of the soul. A corollary notion was that calcification of the pineal caused psychiatric disease, a concept that provided support for those who considered psychotic behaviour to be rampant; modern examination techniques have revealed that all pineal glands become more or less calcified.

Anatomy. The pineal organ is small, weighing little more than 0.1 gram. It lies deep within the brain between the two cerebral hemispheres and above the third ventricle of the spinal column. It has a rich supply of adrenergic nerve fibres that greatly influence its secretions. Microscopically, the gland is composed of pinealocytes (rather typical endocrine cells except for extensions that mingle with those of adjacent cells). Supporting cells that are similar to astrocytes of the brain are interspersed.

Hormones. The pineal gland contains a number of peptides, including GnRH, TRH, and vasotocin, along with a number of important neurotransmitters such as somatostatin, norepinephrine, serotonin, and histamine. The major pineal hormone, however, is melatonin, a derivative of the amino acid tryptophan. Melatonin was first discovered because it lightens amphibian skin, an effect opposite to that of melanocyte-stimulating hormone of the anterior pituitary. Secretion of melatonin is enhanced whenever the sympathetic nervous system is stimulated. Of greater interest, however, is the fact that secretion increases soon after an animal is placed in the dark; the opposite effect takes place immediately upon exposure to light. Its major action, well documented in animals, is to block the secretion of GnRH by the hypothalamus and of gonadotropins by the pituitary. While it was long thought that a decrease in melatonin secretion heralded the onset of puberty, this hypothesis cannot be supported by studies in humans. It is possible that the pineal contains an as yet unidentified hormone that serves that function.

Melatonin

Pineal tumours. Pineal tumours are rare, occurring most often in children and young adults. The most common of these are germ cell tumours (germinomas and teratomas), which arise from embryonic remnants of germ cells. These tumours are malignant and invasive and may be life-threatening. Tumours of pinealocytes also occur and vary in their potential for malignant change.

Pineal tumours may cause headache, vomiting, and seizures due to the increase in pressure within the head that results from the enlarging tumour mass. Endocrinologic effects may also be observed. Some patients may become hypogonadal with regression of secondary sex characteristics, while others may undergo precocious puberty because of secretion of chorionic gonadotropin. Diabetes insipidus is frequently associated and is usually due to tumour invasion of the hypothalamus and posterior pituitary. Invasion of the pituitary stalk may interfere with the ongoing inhibition of prolactin secretion by dopamine from the hypothalamus, resulting in elevated serum prolactin levels, a finding that may lead to a mistaken diagnosis of prolactinoma. Treatment consists of surgical relief of the increased intracranial pressure and X-ray therapy.

HORMONES OF THE INTESTINAL MUCOSA

In 1902, two English physiologists, Sir William M. Bayliss and Ernest H. Starling, placed dilute hydrochloric acid into a segment of a dog's bowel from which the nerve supply had been severed. They then scraped off the bowel lining, boiled it, filtered it, and injected the filtrate into a dog's vein. The injection was followed shortly by a greatly increased secretion of pancreatic juices. They named the unidentified water-soluble material in the filtrate "secretin," and thus was modern endocrinology born. With this discovery emerged the pivotal concept of chemical messages acting at a distant site to regulate bodily functions. Interest in secretin soon waned, however, overshadowed by discoveries in what became the mainstream of endocrinology. It was not until the advent of modern techniques for isolating, characterizing, and measuring protein hormones that interest in the endocrinology of the gastrointestinal tract was revived. It has become clear that the intestinal tract is not only a complex system dedicated to the digestion and absorption of nutrients but also a large endocrine organ that secretes many hormones.

Secretin. Secretin, a polypeptide containing 27 amino acids, is concentrated in the lining of the upper intestine. When hydrochloric acid from the stomach passes into the duodenum, secretin is released into the blood and soon prompts the pancreatic acinar cells to release water and bicarbonate into the pancreatic ducts and from there into the duodenum. By this mechanism, hydrochloric acid, which can be damaging to intestinal lining, is promptly diluted and neutralized by the pancreatic water and bicarbonate. Secretin is used as a stimulator of the pancreas to evaluate exocrine pancreatic functions of patients.

Neutralizing hydrochloric acid

Gastrin. Gastrin is a 17-amino-acid polypeptide that is secreted into the circulation by cells lining the stomach. Gastrin stimulates the secretion of hydrochloric acid and a digestive enzyme, pepsin, into the stomach cavity, while simultaneously increasing the contractions of its distal part. The medical significance of gastrin lies in the fact that there are pancreatic islet cell tumours that secrete large quantities of gastrin or its prohormone, "big gastrin." The affected patient has severe peptic ulcer disease that is unresponsive to the usual forms of treatment. There is often associated diarrhea with bowel movements containing large amounts of fat. Gastrinomas often form part of the syndrome of multiple endocrine neoplasia (MEN I) discussed below. Treatment consists of removing the tumour surgically when feasible or, when not, of cutting the vagus nerve, followed by the administration of a gastric-acid-inhibiting drug such as cimetidine.

Gastric inhibitory polypeptide. Gastric inhibitory polypeptide (GIP) is a hormone secreted by cells of the intestinal mucosa that blocks the secretion of hydrochloric acid into the stomach. It also serves to enhance insulin secretion from the beta cells of the islets of Langerhans so that plasma insulin levels rise after a meal even before the ingested glucose or amino acids enter the blood, an example of an anticipatory hormonal action.

Cholecystokinin. The secretion of cholecystokinin (CCK) is stimulated by the introduction of hydrochloric or fatty acids into the stomach or duodenum. As its name implies, cholecystokinin stimulates the gall bladder to contract and release stored bile into the intestine. Similarly, it stimulates the flow of pancreatic juices. There is interest in the possibility that intestinal hormones, particularly CCK, may induce satiety. According to this hypothesis, after a person eats a meal, the secreted CCK stimulates the satiety centre of the hypothalamus so that the individual "feels full" and stops eating. Because CCK is also known to contract the muscles of the channel leading from the stomach into the duodenum, thus inhibiting gastric emptying, it is possible, however, that people have the feeling of being full simply because of gastric distension.

Vasoactive intestinal polypeptide. Vasoactive intestinal polypeptide (VIP), a 28-amino-acid polypeptide, is secreted by cells throughout the intestinal tract. It acts to change the activity of the intestinal mucosa so that water and electrolytes are secreted rather than absorbed as usual.

Effect on intestinal mucosa

Pancreatic islet cell tumours that secrete excessive amounts of VIP are called VIPomas (Verner-Morrison syndrome). Affected persons have a severe, intractable, debilitating watery diarrhea with an associated loss of large quantities of potassium. If the patient is unable to replace the lost fluids adequately, the resulting dehydration may become life-threatening, leading to use of the term pancreatic cholera. Removal of the tumour and postsurgical chemotherapy has improved the survival rate considerably, even though metastases occur in about one-third of these patients.

Other gastrointestinal hormones also serve as neurotransmitters in the brain, but they have not been found to produce disease. These hormones include katacalcin, caerulein, motilin, neuropeptide Y, and gastrin-releasing peptide (bombesin-like peptide). Glucagon (see above *The pancreas*) and somatostatin (see above *The hypothalamus* and *The pancreas*) also serve as gastrointestinal hormones and brain neurotransmitters, and they also produce rare hyperfunctioning pancreatic tumours.

PROSTAGLANDINS

The prostaglandins (PGs) are a common group of modified fatty acids that are astonishingly diverse in their actions; for the most part, they have a local (paracrine) function. The study of these powerful agents had modest beginnings when, in 1935, a Swedish physiologist and Nobel laureate, Ulf von Euler, and other investigators found that extracts of seminal vesicles or of human semen lowered blood pressure and caused contraction of strips of uterine tissue. Von Euler coined the term prostaglandin because he assumed that the active material came from the prostate gland.

The prostaglandins comprise a group of related cyclic, unsaturated fatty acids that are derived primarily from the 20-carbon, straight-chain, polyunsaturated fatty acid precursor, arachidonic acid. Each prostaglandin differs from the others in subtle changes in chemical structure or side-chain substitutions; these differences are responsible for the different biologic activities of the members of the prostaglandin group.

Arachidonic acid

Arachidonic acid is a key component of the phospholipids, which are themselves integral components of cell membranes. In response to a variety of stimuli, a chain of events is set in motion that results in prostaglandin release (Figure 15).

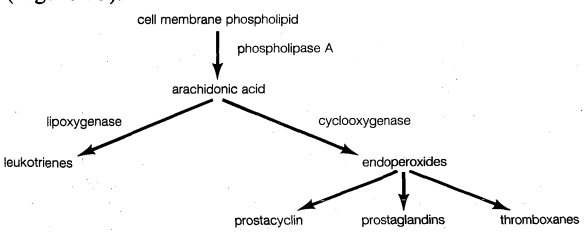


Figure 15: Synthetic pathways of arachidonic acid breakdown from phospholipids of the cell membrane.

The actions of an enzyme, phospholipase A, induce the phospholipids to release the precursor, arachidonic acid. One enzyme, lipoxygenase, catalyzes the synthesis of the leukotrienes. Another enzyme, cyclooxygenase, stimulates the conversion of arachidonic acid to several endoperoxides. The endoperoxides undergo further biosynthesis to the prostaglandins, prostacyclin, and the thromboxanes. (The thromboxanes and prostacyclin are important compounds that have functions in the process of blood coagulation. Leukotrienes, converted from endoperoxides in white blood cells, or leukocytes, are important mediators of the inflammatory process.)

The actions of the prostaglandins are multiple and variable; the same prostaglandin might stimulate a reaction in one tissue and inhibit it in another. Prostaglandin effects are usually manifested locally around the site of prostaglandin synthesis.

When the actions of prostaglandins are stimulatory, they act as intermediaries (necessary elements that elicit a subsequent step along a synthetic or biologic pathway) in the formation of cyclic 3′,5′-adenosine monophosphate (cyclic AMP, cAMP), and thus the final biologic actions of the target cells.

This synthetic pathway begins when tropic hormones (hormones of one endocrine gland that affect the actions of other endocrine structures) are bound to receptors on the surface of the cells of the target organ. These tropic hormones, called first messengers, initiate the prostaglandin-synthesis pathway discussed above, and the increased concentration of prostaglandins around the target organ stimulates the intracellular synthesis of cAMP from adenosine triphosphate (ATP), a process that brings about the biologic action of the target organ. Unaccountably, prostaglandins may inhibit the synthesis of cAMP in some tissues.

Vasodilator effect

Prostaglandins are powerful vasodilators; that is, they relax the muscles in the walls of blood vessels so that the diameters become larger and there is less resistance to the flow. Consequently, the blood pressure falls. Again, the effect can be local. An important example of the vasodilation effect of prostaglandins is found in the kidney, where widespread vasodilation leads to an increase in the flow of blood to the kidney and an increased excretion of salt in the urine. Thromboxanes, on the other hand, are powerful vasoconstrictors in the same setting.

Some diuretics, such as furosemide, probably act by releasing prostaglandins in the kidney. Prostaglandins inhibit the action of vasopressin on the kidney tubules, resulting in enhanced urinary excretion of water. The resultant tendency to dehydration from this enhanced excretion of water leads to local secretion of another kidney prostaglandin that stimulates the secretion of renin (see above *The adrenal cortex: Aldosterone*). Renin stimulates the production of aldosterone, which has the affect of conserving sodium and water, thus combating the dehydration and elevating the depressed blood pressure.

Although prostaglandins were first detected in semen, no biologic role for them has been defined in the male reproductive system. This is not true, however, for females. It has been shown that prostaglandins mediate the control of GnRH over LH secretion, modulate ovulation, and stimulate uterine muscle contraction. Discovery of this last property has led to the successful treatment of menstrual cramps (dysmenorrhea) through the use of inhibitors of prostaglandin synthesis, such as ibuprofen. Prostaglandins also play a role in inducing labour in pregnant women at term or in inducing therapeutic abortions.

The process of clot formation begins with an aggregation of blood platelets. This process is strongly stimulated by thromboxanes and inhibited by prostacyclin. Prostacyclin is synthesized in the walls of blood vessels and serves the physiological function of preventing needless clotting. Thromboxanes, on the other hand, are synthesized within the platelets themselves and are released. The platelets adhere to one another and to blood vessel walls. Through prostaglandin and thromboxane mechanisms, clotting is prevented when it is unnecessary and takes place when it is necessary. Platelets adhere in arteries that are affected by the process of atherosclerosis; they form plaques along the interior surface of the vessel wall. This type of platelet aggregation and clotting leads to blocking (occlusion) of the vessel wall, the most common cause of heart attack (coronary artery occlusion). This biologic insight has led to the widespread recommendation that those at risk for a coronary occlusion take aspirin, an inhibitor of the enzyme cyclooxygenase, daily as a preventive measure.

Role in inflammation

Prostaglandins also play a pivotal role in inflammation, a process characterized by the ancient Romans as consisting of redness (*rubor*), heat (*calor*), pain (*dolor*), and swelling (*tumor*). These changes are due to a local dilation of blood vessels that permits increased blood flow to the affected area. The blood vessels become more permeable, leading to the escape of infection-fighting fluid and white blood cells from the blood into the surrounding tissues. These changes are mediated by prostaglandins, particularly the subgroup called leukotrienes. Thus, effective treatment to suppress inflammation in inflammatory but noninfectious diseases, such as rheumatoid arthritis, is to treat the patient with inhibitors of prostaglandin synthesis, such as aspirin. Similarly, the pain and fever of other disseminated inflammations can be alleviated by these nonsteroidal anti-inflammatory drugs.

Another crucial mechanism of the body that protects it from invasion by bacteria, viruses, or other noxious agents is known as the immune response. It begins when a foreign substance is ingested by a mobile, scavenging, white blood cell, called a macrophage. The macrophage interacts with a special white blood cell called a T-lymphocyte (T cell), which in turn activates B-lymphocytes (B cells or plasma cells). The result is that the B cell elaborates and secretes specific proteins (antibodies) that are designed to make the ingested foreign invader more susceptible to attack and ingestion by other white blood cells.

In cellular immune response, T cells become activated at the site of damage and release proteins called lymphokines, which attract macrophages to the local area and stimulate them to ingest the offending agents. Prostaglandins generally attenuate the immune response by inhibiting both T cell and B cell activity, but some prostaglandins, particularly the leukotrienes, enhance inflammatory responses.

The understanding of the immune response marks a ma-

Anaphylactic reactions

jor advance in medicine since aberrations in this response cause hypersensitivity (anaphylactic) reactions, allergies, and autoimmune diseases. Examples include harmful reactions to drugs such as penicillin; hay fever; bronchial asthma; rheumatoid arthritis; Graves' disease; and autoimmune endocrine deficiency diseases. Prostaglandins play important roles in the genesis of these disorders, an awareness that has led to the development of a number of powerful inhibitors of prostaglandin synthesis for use in treatment.

The functioning of the digestive tract is also influenced by prostaglandins. Depending on the setting, various prostaglandins may either enhance or inhibit the contraction of the smooth muscles of the intestinal walls. They are also powerful inhibitors of stomach secretions, perhaps because they inhibit the secretion of the stomach hormone gastrin, which stimulates gastric secretion. It is not surprising, then, that drugs, like aspirin, which inhibit prostaglandin synthesis may lead to peptic ulcers. Prostaglandin action on the digestive tract may cause a severe watery diarrhea and may mediate the effects of vasoactive intestinal polypeptide (VIP) in the Verner-Morrison syndrome (see above *Hormones of the intestinal tract*), as well as the effects of cholera toxin.

Prostaglandins induce several effects on endocrine function. Perhaps of greatest importance is the ability of prostaglandins to stimulate the resorption of bone in diseases such as rheumatoid arthritis and to cause hypercalcemia, particularly in patients harbouring malignant tumours.

The therapeutic applications of the prostaglandins and of the drugs that inhibit prostaglandin synthesis are listed in Table 3. The drugs fall into two categories. In the first are agents like hydrocortisone and its synthetic derivatives, such as prednisone, which stabilize cell membranes and, in large doses, block the liberation of arachidonic acid. In the second are drugs that block the action of the enzyme cyclooxygenase. Among these are aspirin, acetaminophen, indomethacin, and ibuprofen.

Table 3: Therapeutic Applications of the Prostaglandins*

prostaglandins	PG synthesis inhibition
Current	
Midtrimester abortion	Rheumatoid arthritis
Peripheral vascular disease	Fever and headache
Hemodialysis	Bartter's syndrome
Induction of labour	Patent ductus arteriosus
Potential	
Hypertension	Hypercalcemia of malignant disease
Congestive heart failure	Periodontal inflammation
Infertility	Cholera and certain diarrheal states
Coronary and deep thrombosis	Burns
Peptic ulceration	Lupus erythematosus
Gastric hyperacidity	Glaucoma
Bronchial asthma (PGE)	Migraine headache
Nasal congestion	Bronchial asthma (leukotriene)

*From J.D. Wilson and D.W. Foster (eds.), *Williams Textbook of Endocrinology*, 7th ed., Philadelphia, W.B. Saunders Co., 1985. Reprinted by permission.

ECTOPIC HORMONE AND POLYGLANDULAR DISORDERS

Tumour development

In discussing general characteristics of endocrine hyperfunction above it was indicated that the cells of endocrine glands, following long-term stimulation, increase in size (hypertrophy) and number (hyperplasia). If the stimulation persists, these cells may be transformed into a tumour, which may be either benign or malignant.

For reasons that have aroused much speculation but remain poorly understood, these changes may occur in more than one endocrine gland, simultaneously or consecutively, even though the embryonic origin of the cells of the other endocrine glands that are involved may be different, and even though this propensity to tumour formation is confined to endocrine glands only (multiple endocrine neoplasia). Similarly, for equally obscure reasons, multiple endocrine glands may be attacked by autoantibodies with the result that the patient is afflicted with multiple hormonal deficiencies (multiple endocrine deficiency syndromes). Finally, there have emerged syndromes due to excessive amounts of hormones produced by tumours of tissues that do not ordinarily produce hormones at all (ectopic hormone production). This transformation initially was thought to occur when the tumours activated genes that generated these hormones, and that ordinarily were repressed in the cell of the nonendocrine tissue. Recent evidence has revealed that most tissues synthesize small amounts of most hormones and, indeed, other substances that have no hormonal activity, so that the change that occurs following tumour formation is a quantitative rather than a qualitative one.

Multiple endocrine neoplasia. Multiple endocrine neoplasias (MEN) are hereditary disorders usually occurring in an autosomal dominant genetic distribution (*i.e.*, the defect is not tied to the sex of the individual and statistically, one-half of the children of an affected person will also be affected) so that families are heavily sprinkled with affected individuals. There are several defined patterns of glandular involvement which usually, but not always, "breed true" in that the clusters of glandular involvement follow the same groupings from one family member to another. Studies of distribution in humans are necessarily incomplete because the endocrine tumours do not appear simultaneously. Thus, a patient who may appear to have an incomplete expression of one of these inherited syndromes when first examined may later develop the full clinical picture.

Principal glands involved

The first described and the most frequently occurring of these unusual disorders is multiple endocrine neoplasia type I (MEN I). The principal glands involved in this syndrome are the parathyroids, the pancreatic islets, and the anterior pituitary. All four parathyroid glands are involved either by hyperplasia alone or a mixture of hyperplasia and adenomas. The symptoms of hyperparathyroidism may be mild and are often overshadowed by the disabling problems engendered by the islet cell tumours. Two-thirds of patients with MEN I develop gastrinomas with severe, intractable peptic ulcers. Insulinomas with severe hypoglycemia also occur frequently, and in some patients both tumours arise. The pituitary manifestations are most frequently those of prolactinoma or acromegaly (see above *The anterior pituitary*). Involvement of the adrenal cortex and the thyroid glands may occur, but it is possible that these aberrations are coincidental rather than an integral part of the hereditary disease. Treatment consists of attacks on individual hyperfunctioning glands as they appear; however, in contrast to sporadic cases, it is important to counsel families to have all members screened for evidence of MEN I because early treatment is more effective and less risky.

Multiple endocrine neoplasia type II (MEN II) is composed of another distinct constellation of glandular involvements: a medullary carcinoma of the thyroid; pheochromocytoma, usually bilateral; and, again, hyperparathyroidism. Medullary carcinomas of the thyroid arise from the parafollicular C cells (see above *The thyroid gland*), which secrete calcitonin (see above *The parathyroid glands*). Medullary thyroid carcinoma occurs in all affected families except those who, by screening techniques, are detected at an early stage when C cell hyperplasia has not yet been transformed into a carcinoma. Medullary thyroid carcinoma is an example of ectopic hormone production in that these tumours may elaborate excessive quantities not only of the expected hormone, calcitonin, but also ectopically of other bioactive substances, including corticotropin, prostaglandins, serotonin, and the neurotransmitter substance P. While only a small minority of all patients harbouring pheochromocytoma have MEN II, when these tumours do occur in both adrenal glands, the likelihood is much greater that MEN II is present. As in the case of C cells, adrenal medullary hyperplasia precedes the development of true tumour formation. The high blood pressure and other symptoms characteristic of pheochromocytoma have been described previously (see above *The adrenal medulla*), and, again, parathyroid hyperplasia occurs more frequently than parathyroid tumours in this syndrome. Early screening of family members is strongly recommended, and treatment does not

differ from that applied to patients with a single hyperfunctioning endocrine gland.

MEN IIB

A variant of MEN II is termed MEN IIB, or MEN III. Patients with this disease also suffer from medullary thyroid carcinoma and pheochromocytoma, but they differ in that hyperparathyroidism rarely occurs, and affected family members uniformly develop mucosal neuromas. These are nerve tumours, usually benign, involving the lips and linings of the mouth, nose, and throat. They may be recognized at birth or in early childhood as "bumpy lips," and these neuromas may be scattered throughout the gastrointestinal tract, causing constipation and, less frequently, vomiting and difficulty in swallowing.

On rare occasions, some patients with multiple endocrine neoplasia do not fit established patterns. Some of these aberrations may be explained by coincidence, but others seem to represent true MEN of a mixed type, for example, pheochromocytoma associated with pancreatic islet cell tumours.

Multiple endocrine deficiency syndromes. In multiple endocrine deficiency syndromes, affected families have some or all of a bewildering array of ailments shown in Table 4. Investigators have found it convenient to divide these deficiency diseases into two types, although, from inspection of the Table, it can be seen that there is considerable overlap. Type II is inherited in an autosomal-dominant pattern, while type I is thought to be due to autosomal-recessive inheritance. What is inherited in both types is the propensity to develop circulating autoantibodies directed against, and destroying, one or more of the tissues listed in Table 4.

Table 4: Autoimmune Polyglandular Syndromes*

type II	type I
Hyperthyroidism	Hypoparathyroidism
Hypothyroidism	Yeast (*Candida*) infection of the skin and mucous membranes
Diabetes, type I	Adrenal insufficiency
Adrenal insufficiency	Hepatitis
Myasthenia gravis	Malabsorption
Celiac disease	Vitiligo
Hypogonadism	Pernicious anemia
Pernicious anemia	Alopecia
Vitiligo (skin depigmentation)	Hypothyroidism
Alopecia	

*Adapted from J.D. Wilson and D.W. Foster (eds.), *Williams Textbook of Endocrinology*, 7th ed., Philadelphia, W.B. Saunders Co., 1985. Reprinted by permission.

These autoantibodies (see above *The human endocrine system: Endocrine dysfunction*) may be detected in the blood many years before the discernible disease appears. The apparently paradoxical appearance of hyperthyroidism in type II results from the development of circulating thyroid-stimulating autoantibodies. Type II diseases are distinguishable from type I in that multiple generations are affected, the highest incidence occurring between the ages of 20 and 60, and that affected individuals are not afflicted with mucocutaneous candidiasis (a fungal infection of the mucous membranes and skin). Type I, in contrast, has its onset in infancy or childhood, it is commonly associated with candidiasis, and it is characterized by the appearance of the disease in siblings but without transmission from one generation to the next. Treatment is directed toward each individual abnormality.

Onset of MEN syndromes

Ectopic hormone production. Previous views that it is rare for excessive quantities of hormones to be secreted by tumours of nonendocrine origin have been supplanted by demonstrations that ectopic hormone production is indeed quite common (Table 5). Ectopic corticotropin production, the most common of these syndromes, is most frequently associated with carcinoma of the lung, carcinoma of the thymus, or islet cell tumour; however, it may also occur in association with a long list of other neoplasms, including pheochromocytoma, bronchial adenoma, medullary thyroid carcinoma, and carcinomas of the ovary, prostate, breasts, kidney, testes, gallbladder, and even of the appendix. Patients usually have the intense pigmentation and severe depletion of potassium that is characteristic of overproduction of ACTH and of mineralocorticoids. In addition the excessive tissue breakdown characteristic of Cushing's syndrome is added to the debilitating effects of the cancer itself. Treatment ordinarily involves removal or destruction of the cancer, but occasionally, when the tumour cannot be completely removed, an attack on the overactive adrenals, either with drugs or by surgical removal, is warranted.

Table 5: Hormones and Hormone Precursors Reported to be Produced by Neoplasms*

ACTH, lipotropin, and pro-opiomelanocortin
Corticotropin-releasing hormone
Chorionic gonadotropin and its subunits (α and β)
Vasopressin
Growth factors (*e.g.*, IGF)
Parathyroid hormonelike materials
Osteoclast-activating factor
Erythropoietin
Eosinophilopoietin
Growth hormone
Growth hormone-releasing hormone
Prolactin
Gastrin
Gastrin-releasing peptide (and bombesin)
Secretin
Glucagon
Calcitonin
Renin
Vasoactive intestinal peptide
Somatostatin
Hypophosphatemia-producing factor
Prostaglandins
Estrone and estradiol

*From J.D. Wilson and D.W. Foster (eds.), *Williams Textbook of Endocrinology*, 7th ed., Philadelphia, W.B. Saunders Co., 1985. Reprinted by permission.

Chorionic gonadotropin

The synthesis of chorionic gonadotropin (a hormone produced by the placenta that stimulates the gonads) originally was thought to be confined to one of the membranes covering the fetus. Recent, more sensitive testing, however, has revealed that at least one segment of this hormone is synthesized in almost all tissues. Chorionic gonadotropin is a glycoprotein similar to TSH, LH, and FSH in that it contains both alpha and beta chains. It is likely that it is a fragment of the beta chain that is found in tumour tissues. As much as 13 percent of all carcinomas are associated with increased circulating levels of chorionic gonadotropin-like material. Affected patients may have no symptoms or may have symptoms similar to those produced by excessive LH secretion.

There are numerous other manifestations of ectopic hormone production. They include the secretion of bioactive materials that result in hypoglycemia, hypercalcemia, hypocalcemia, and the inappropriate secretion of vasopressin (see above *The posterior pituitary*), growth hormone, and growth-hormone-releasing hormone. As discussed above, treatment is aimed at ablating or reducing the activity of the offending tumour or mitigating the effects of the hormone produced in excess.

ENDOCRINE CHANGES WITH AGING

Because the endocrine glands play pivotal roles both in reproduction and in development, it seems plausible to extend the role of the endocrine system to account for the progressive bodily changes that occur with aging (senescence). Indeed, for a time, an "endocrine theory of aging" enjoyed wide popularity among scientists. Early in the 20th century, the possibility that aging could be deferred and virility restored by the injection of crude extracts of monkey glands attracted a good deal of attention. Upon closer scrutiny, however, it has become clear that the endocrine glands weather the ravages of age quite well and, in a number of instances, tend to mitigate its effects. (For a discussion of the aging process, see GROWTH AND DEVELOPMENT, BIOLOGICAL.)

The menopause. The most striking change with age is that of the menopause (see above *The ovary*). Estrogens are produced by granulosa cells and cells of the stroma, which line the egg-containing ovarian follicles. Because the number of these follicles in the ovaries is limited, their depletion with age makes inevitable the reduction in estrogen

Reduced estrogen levels

secretion, which, in endocrinologic terms, defines the onset of the menopause. The low circulating estrogen levels reduce hypothalamic and pituitary inhibition of GnRH, LH, and FSH secretion so that circulating levels of these hormones undergo a striking and sustained elevation; a three- to fourfold increase above premenopausal values is found in women above the age of 60. Prolactin secretion also increases. Clearly, in normal postmenopausal women, while the ovaries have "failed" to a large degree, the hypothalamus and pituitary have not.

The testis. Reduction in the number of androgen-secreting Leydig cells leads to a tendency toward a decrease in serum testosterone levels, which are compensated for by an increase in gonadotropin secretion. The result is that the healthy, aging male maintains androgen synthesis and secretion at or near normal levels and may father children despite a greatly advanced age. (For further discussion of changes in the gonadal axis in males with age, see above *The testis.*)

Thyroid and adrenal function. Changes in thyroid function with age are subtle and have limited clinical significance. Circulating levels of the thyroid hormone T_4 remain normal while those of the thyroid hormone T_3 tend to decrease. There appears to be reduced responsiveness of TSH-secreting cells to stimulation with thyrotropin-releasing hormone. Some slowing of the metabolic rate may serve well the "weary bones" and tissues of the healthy aged. Similarly, the hypothalamic-pituitary-adrenocortical axis undergoes minor changes but remains intact with advancing years. Plasma cortisol levels remain essentially unchanged. Aldosterone secretion decreases as do plasma renin concentrations, but the healthy elderly are able to maintain normal balances of fluids and electrolytes (see above *The adrenal cortex*).

Growth hormone, parathyroid, and antidiuretic hormones. Growth hormone secretion decreases variably with age. In some healthy elderly persons it is moderately reduced as compared to young adults, and in some otherwise apparently healthy aged individuals there seem to

Growth hormone

be deficient responses in growth hormone secretion. It is possible, then, that a subpopulation of the aging population may benefit from growth hormone treatment. Serum parathyroid levels seem to rise with age, a change that may serve to maintain normal serum calcium levels. Similarly, the secretion of antidiuretic hormone (vasopressin) tends to be elevated and hyperresponsive. This may occur in response to an increasing difficulty of the aging kidneys to prevent inordinate excretion of water.

The pancreatic islets. It has been well documented that blood sugar levels, while normal in the fasting state, respond to the ingestion of glucose with increments proportional to the age of the subject; that is, the older the healthy subject, the higher the maximal increase in blood glucose after glucose ingestion. The accompanying increase in levels of serum insulin, although appreciable, is clearly not enough to maintain the glucose levels in the range found in healthy young adults. Whether these changes should be viewed as abnormal or whether they merely reflect modifications appropriate to the aging process remains a matter of debate.

In summary, endocrine changes in healthy aging individuals do not account for the aging process. There is evidence that, with aging, there is a progressive loss in the numbers of hormonal tissue receptors, and, more often than not, there is an appropriate increase in hormone secretion to maintain a healthy homeostatic balance. The case of the failing ovary excepted, the endocrine glands generally sustain their major function of supporting a state of health in the face of declining tissue and organ function until such time as the whole organism falters and decrepitude ensues. (T.B.S.)

BIBLIOGRAPHY

General works: A comprehensive historical and biographical survey is provided by VICTOR CORNELIUS MEDVEI, *A History of Endocrinology* (1982). Comprehensive standard texts include JEAN D. WILSON and DANIEL W. FOSTER (eds.), *Williams Textbook of Endocrinology,* 7th ed. (1985); PHILIP FELIG *et al.* (eds.), *Endocrinology and Metabolism,* 2nd ed. (1987); LESLIE J. DEGROOT *et al.* (eds.), *Endocrinology,* 3 vol. (1979); and FRANCIS S. GREENSPAN and PETER H. FORSHAM (eds.), *Basic & Clinical Endocrinology,* 2nd ed. (1986). For modern research in the field, see *Recent Progress in Hormone Research: Proceedings of the Laurentian Hormone Conference* (irregular); and *Current Therapy in Endocrinology and Metabolism* (biennial). PETER H. WISE, *Endocrinology* (1986), is a useful atlas.

Briefer coverage is provided in JAY TEPPERMAN and HELEN M. TEPPERMAN, *Metabolic and Endocrine Physiology: An Introductory Text,* 5th ed. (1987); ROBERT VOLPÉ (ed.), *Autoimmunity and Endocrine Disease* (1985); C. DONNELL TURNER and JOSEPH T. BAGNARA, *General Endocrinology,* 6th ed. (1976); C.R. KANNAN, *Essential Endocrinology: A Primer for Nonspecialists* (1986); E.D. WILLIAMS (ed.), *Current Endocrine Concepts* (1982); BRIAN K. FOLLETT, SUSUMU ISHII, and ASHA CHANDOLA (eds.), *The Endocrine System and the Environment* (1985); and T.S. DANOWSKI, *Outline of Endocrine Gland Syndromes,* 3rd ed. (1976). A survey of medical literature can be found in *The Year Book of Endocrinology.*

Glands and hormones: SEYMOUR REICHLIN, ROSS J. BALDESSARINI, and JOSEPH B. MARTIN (eds.), *The Hypothalamus* (1978); CHOH HAO LI (ed.), *Hypothalamus Hormones* (1979); PETER J. MORGANE and JAAK PANKSEPP (eds.), *Handbook of the Hypothalamus,* 3 vol. in 4 (1979–81); AJAY S. BHATNAGAR (ed.), *The Anterior Pituitary Gland* (1983); PETER H. BAYLIS and PAUL L. PADFIELD (eds.), *The Posterior Pituitary: Hormone Secretion in Health and Disease* (1985); GEORGE T. TINDALL, DANIEL L. BARROW, and JOSEPH B. MARTIN, *Disorders of the Pituitary* (1986); SIDNEY H. INGBAR and LEWIS E. BRAVERMAN (eds.), *Werner's The Thyroid: A Fundamental and Clinical Text,* 5th ed. (1986); PATRICK J. MULROW (ed.), *The Adrenal Gland* (1986); RUSSEL J. REITER (ed.), *The Pineal Gland* (1984); G.M. BROWN and S.D. WAINWRIGHT (eds.), *The Pineal Gland: Endocrine Aspects* (1985); and R.J. WURTMAN and F. WALDHAUSER (eds.), *Melatonin in Humans* (1986).

Gynecological and reproductive endocrinology: SAMUEL S.C. YEN and ROBERT B. JAFFE, *Reproductive Endocrinology: Physiology, Pathophysiology, and Clinical Management,* 2nd ed. (1986); PHILIP RHODES, *Reproductive Physiology* (1969); DANIEL R. MISHELL, JR., and VAL DAVAJAN (eds.), *Infertility, Contraception, & Reproductive Endocrinology,* 2nd ed. (1986); KYOICHIRO OCHIAI *et al.* (eds.), *Endocrine Correlates of Reproduction* (1984); JOHN E. TYSON (ed.), *Neuroendocrinology of Reproduction* (1978); and EUGENE D. ALBRECHT and GERALD J. PEPE (eds.), *Perinatal Endocrinology* (1985).

Diabetes mellitus and hypoglycemia: SYDNEY S. LAZARUS and BRUNO W. VOLK, *The Pancreas in Human and Experimental Diabetes* (1962); BRUNO W. VOLK and EDWARD R. ARGUILLA (eds.), *The Diabetic Pancreas,* 2nd ed. (1985); ELLIOTT P. JOSLIN, *Joslin's Diabetes Mellitus,* 12th ed., edited by ALEXANDER MARBLE *et al.* (1985); MAYER B. DAVIDSON, *Diabetes Mellitus: Diagnosis and Treatment,* 2nd ed. (1986); BERNARD N. BRODOFF and SHELDON J. BLEICHER (eds.), *Diabetes Mellitus and Obesity* (1982); and DOROTHY REYCROFT HOLLINGSWORTH, *Pregnancy, Diabetes, and Birth* (1984). For current research, see *Diabetes* (monthly).

Neuroendocrinology: BERNARD T. DONOVAN, *Hormones and Human Behaviour* (1985); KENNETH W. MCKERNS and VLADIMIR PANTIĆ (eds.), *Neuroendocrine Correlates of Stress* (1985); NANDKUMAR S. SHAH and ALEXANDER G. DONALD (eds.), *Psychoendocrine Dysfunction* (1984); DEREK GUPTA, PATRIZIA BORRELLI, and ANDREA ATTANASIO (eds.), *Paediatric Neuroendocrinology* (1985); JOSEPH B. MARTIN and SEYMOUR REICHLIN, *Clinical Neuroendocrinology,* 2nd ed. (1987); and JOSEPH MEITES (ed.), *Neuroendocrinology of Aging* (1983). For current research, see *Frontiers in Endocrinology* (irregular).

Comparative endocrinology: DAVID O. NORRIS, *Vertebrate Endocrinology,* 2nd ed. (1985); ARI VAN TIENHOVEN, *Reproductive Physiology of Vertebrates,* 2nd ed. (1983); KENNETH C. HIGHNAM and LEONARD HILL, *The Comparative Endocrinology of the Invertebrates,* 2nd ed. (1977); GEOFFREY W. BENNETT and SAFFRON A. WHITEHEAD, *Mammalian Neuroendocrinology* (1983); E.J.W. BARRINGTON and C. BARKER JØRGENSEN (eds.), *Perspectives in Endocrinology: Hormones in the Lives of Lower Vertebrates* (1968); and AUBREY GORBMAN *et al., Comparative Endocrinology* (1983).

(T.B.S.)

Energy Conversion

Over the centuries a wide array of devices and systems has been developed for converting energy from forms provided by nature to those most useful to society. Some of these energy converters are quite simple. The early windmills, for example, transformed the kinetic energy of wind into mechanical energy for pumping water and grinding grain. Other energy-conversion systems are decidedly more complex, particularly those that take raw energy from fossil fuels and nuclear fuels to generate electrical power. Systems of this kind require multiple steps or processes in which energy undergoes a whole series of transformations through various intermediate forms.

Many of the energy converters widely used today involve the transformation of thermal energy into electrical energy. The efficiency of such systems is, however, subject to fundamental limitations, as dictated by the laws of thermodynamics and other scientific principles. In recent years, considerable attention has been devoted to certain direct energy-conversion devices, notably solar cells and fuel cells, that bypass the intermediate step of conversion to heat energy in electrical power generation.

This article traces the development of energy-conversion technology, highlighting not only conventional systems but also alternative and experimental converters with considerable potential. It delineates their distinctive features, basic principles of operation, major types, and key applications. For a discussion of the laws of thermodynamics and their impact on system design and performance, see THERMODYNAMICS, PRINCIPLES OF.

For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 112, 123, 124, 127, 711, and 721, and the *Index.*

The article is divided into the following sections:

FUNDAMENTALS OF ENERGY CONVERSION

General considerations

Definition of energy

Energy is usually and most simply defined as the equivalent of or capacity for doing work. The word itself is derived from the Greek *energeia: en,* "in"; *ergon,* "work." Energy can either be associated with a material body, as in a coiled spring or a moving object, or it can be independent of matter, as light and other electromagnetic radiation traversing a vacuum. The energy in a system may be only partly available for use. The dimensions of energy are those of work, which, in classical mechanics, is defined formally as the product of mass (m) and the square of the ratio of length (l) to time (t): ml^2/t^2. This means that the greater the mass or the distance through which it is moved or the less the time taken to move the mass, the greater will be the work done, or the greater the energy expended.

DEVELOPMENT OF THE CONCEPT OF ENERGY

The term energy was not applied as a measure of the ability to do work until rather late in the development of the science of mechanics. Indeed, the development of classical mechanics may be carried out without recourse to the concept of energy. The idea of energy, however, goes back at least to Galileo in the 17th century. He recognized that, when a weight is lifted with a pulley system, the force applied multiplied by the distance through which that force must be applied (a product called, by definition, the work) remains constant even though either factor may vary. The concept of vis viva, or living force, a quantity directly proportional to the product of the mass and the square of the velocity, was introduced in the 17th century. In the 19th century the term energy was applied to the concept of the vis viva.

Isaac Newton's first law of motion recognizes force as being associated with the acceleration of a mass. It is almost inevitable that the integrated effect of the force acting on the mass would then be of interest. Of course, there are two kinds of integral of the effect of the force acting on the mass that can be defined. One is the integral of the force acting along the line of action of the force, or the spatial integral of the force; the other is the integral of the force over the time of its action on the mass, or the temporal integral.

Evaluation of the spatial integral leads to a quantity that is now taken to represent the change in kinetic energy of the mass resulting from the action of the force and is just one-half the vis viva. On the other hand, the temporal integration leads to the evaluation of the change in momentum of the mass resulting from the action of the force. For some time there was debate as to which integration led to the proper measure of force, the German philosopher-scientist Gottfried Wilhelm Leibniz arguing for the spatial integral as the only true measure, while earlier the French philosopher and mathematician René Descartes had defended the temporal integral. Eventually, in the 18th cen-

tury, the physicist Jean d'Alembert of France showed the legitimacy of both approaches to measuring the effect of a force acting on a mass and that the controversy was one of nomenclature only.

To recapitulate, force is associated with the acceleration of a mass; kinetic energy, or energy resulting from motion, is the result of the spatial integration of a force acting on a mass; momentum is the result of the temporal integration of the force acting on a mass; and energy is a measure of the capacity to do work. It might be added that power is defined as the time rate at which energy is transferred (to a mass as a force acts on it, or through transmission lines from the electrical generator to the consumer).

Conservation of energy (see below) was independently recognized by many scientists in the first half of the 19th century. The conservation of energy as kinetic, potential, and elastic energy in a closed system under the assumption of no friction has proved to be a valid and useful tool. Further, upon closer inspection, the friction, which serves as the limitation on classical mechanics, is found to express itself in the generation of heat, whether at the contact surfaces of a block sliding on a plane or in the bulk of a fluid in which a paddle is turning or any of the other expressions of "friction." Heat was identified as a form of energy by Hermann von Helmholtz of Germany and James Prescott Joule of England during the 1840s. Joule also proved experimentally the relationship between mechanical and heat energy at this time. As more detailed descriptions of the various processes in nature became necessary, the approach was to seek rational theories or models for the processes that allow a quantitative measure of the energy change in the process and then to include it and its attendant energy balance within the system of interest, subject to the overall need for the conservation of energy. This approach has worked for the chemical energy in the molecules of fuel and oxidizer liberated by their burning in an engine to produce heat energy that subsequently is converted to mechanical energy to run a machine; it has also worked for the conversion of nuclear mass into energy in the nuclear fusion and nuclear fission processes.

Heat as a form of energy

ENERGY CONSERVATION AND TRANSFORMATION

The concept of energy conservation. A fundamental law that has been observed to hold for all natural phenomena requires the conservation of energy—*i.e.*, that the total energy does not change in all the many changes that occur in nature. The conservation of energy is not a description of any process going on in nature, but rather it is a statement that the quantity called energy remains constant regardless of when it is evaluated or what processes—possibly including transformations of energy from one form into another—go on between successive evaluations.

The law of conservation of energy is applied not only to nature as a whole but to closed or isolated systems within nature as well. Thus, if the boundaries of a system can be defined in such a way that no energy is either added to or removed from the system, then energy must be conserved within that system regardless of the details of the processes going on inside the system boundaries. A corollary of this closed-system statement is that whenever the energy of a system as determined in two successive evaluations is not the same, the difference is a measure of the quantity of energy that has been either added to or removed from the system in the time interval elapsing between the two evaluations.

Possible forms of energy within a system

Energy can exist in many forms within a system and may be converted from one form to another within the constraint of the conservation law. These different forms include gravitational, kinetic, thermal, elastic, electrical, chemical, radiant, nuclear, and mass energy. It is the universal applicability of the concept of energy, as well as the completeness of the law of its conservation within different forms, that makes it so attractive and useful.

Transformation of energy. *An ideal system.* A simple example of a system in which energy is being converted from one form to another is provided in the tossing of a ball with mass m into the air. When the ball is thrown vertically from the ground, its speed and thus its kinetic energy decreases steadily until it comes to rest momentarily at its highest point. It then reverses itself, and its speed and kinetic energy increase steadily as it returns to the ground. The kinetic energy E_k of the ball at the instant it left the ground (point 1) was half the product of the mass and the square of the velocity, or $\frac{1}{2}mv_1^2$, and decreased steadily to zero at the highest point (point 2). As the ball rose in the air, it gained gravitational potential energy E_p. Potential in this sense does not mean that the energy is not real but rather that it is stored in some latent form and can be drawn upon to do work. Gravitational potential energy is energy that is stored in a body by virtue of its position in the gravitational field. Gravitational potential energy of a mass m is observed to be given by the product of the mass, the height h attained relative to some reference height, and the acceleration g of a body resulting from the Earth's gravity pulling on it, or mgh. At the instant the ball left the ground at height h_1 its potential energy E_{p1} is mgh_1. At its highest point, its potential energy E_{p2} is mgh_2. Applying the law of conservation of energy and assuming no friction in the air, these add up to form the following equations:

Potential and kinetic energy

$$E_{k1} + E_{p1} = E_{k2} + E_{p2}$$

or

$$\frac{1}{2}mv_1^2 + mgh_1 = 0 + mgh_2.$$

In this idealized example the kinetic energy of the ball at ground level is converted into work in raising the ball to h_2 where its gravitational potential energy has been increased by $mg(h_2 - h_1)$. As the ball falls back to the ground level h_1, this gravitational potential energy is converted back into kinetic energy and its total energy at h_1 again is $\frac{1}{2}mv_1^2 + mgh_1$. In this chain of events the kinetic energy of the ball is unchanged at h_1; thus the work done on the ball by the force of gravity acting on it in this cycle of events is zero. This system is said to be a conservative one.

Varying degrees of conversion in real systems. Although the total amount of energy in an isolated system remains unchanged, there may be a great difference in the quality of different forms of energy. Many forms of energy, in theory, can be transformed completely into work or into other forms of energy. This is true for mechanical energy and electrical energy. The random motions of constituent parts of a material associated with thermal energy, however, represent energy that is not available completely for conversion into directed energy.

The French engineer Sadi Carnot described (in 1824) a theoretical power cycle of maximum efficiency for converting thermal into mechanical energy. He demonstrated that this efficiency is determined by the magnitude of the temperatures at which heat energy is added and waste heat is given off during the cycle. A practical engine operating on the Carnot cycle has never been devised, but the Carnot cycle determines the maximum efficiency of thermal energy conversion into any form of directed energy. The Carnot criterion renders 100 percent efficiency impossible for all heat engines. In effect, it constitutes the basis for what is now the second law of thermodynamics.

The so-called Carnot efficiency

(R.L.Se./C.R.R./Ed.)

History of energy-conversion technology

EARLY ATTEMPTS TO HARNESS NATURAL FORMS OF ENERGY

Early humans first made controlled use of an external, nonanimal energy source when they discovered how to use fire. Burning dried plant matter (primarily wood) and animal waste, they employed the energy from this biomass for heating and cooking. The generation of mechanical energy to supplant human or animal power came very much later—only about 2,000 years ago—with the development of simple devices to harness the energy of flowing water and of wind.

Waterwheels. The earliest machines were waterwheels, first used for grinding grain. They were subsequently adopted to drive sawmills and pumps, to provide the bellows action for furnaces and forges, to drive tilt hammers or trip-hammers for forging iron, and to provide direct me-

Diverse applications

chanical power for textile mills. Until the development of steam power during the Industrial Revolution at the end of the 18th century, waterwheels were the primary means of mechanical power production, rivaled only occasionally by windmills. Thus, many industrial towns, especially in early America, sprang up at locations where water flow could be assured all year.

The oldest reference to a water mill dates to about 85 BC, appearing in a poem by an early Greek writer celebrating the liberation from toil of the young women who operated the querns (primitive hand mills) for grinding corn. According to the Greek geographer Strabo, King Mithradates VI of Pontus in Asia used a hydraulic machine, presumably a water mill, by about 65 BC.

Early vertical-shaft water mills drove querns where the wheel, containing radial vanes or paddles and rotating in a horizontal plane, could be lowered into the stream. The vertical shaft was connected through a hole in the stationary grindstone to the upper, or rotating, stone. The device spread rapidly from Greece to other parts of the world, because it was easy to build and maintain and could operate in any fast-flowing stream. It was known in China by the 1st century AD, was used throughout Europe by the end of the 3rd century, and had reached Japan by the year 610. Users learned early that performance could be improved with a millrace and a chute that would direct the water to one side of the wheel.

A horizontal-shaft water mill was first described by the Roman architect and engineer Vitruvius about 27 BC. It consisted of an undershot waterwheel in which water enters below the centre of the wheel and is guided by a millrace and chute. The waterwheel was coupled with a right-angle gear drive to a vertical-shaft grinding wheel. This type of mill became popular throughout the Roman Empire, notably in Gaul, after the advent of Christianity led to the freeing of slaves and the resultant need for an alternative source of power. Early large waterwheels, which measured about 1.8 metres (six feet) in diameter, are estimated to have produced about three horsepower, the largest amount of power produced by any machine of the time. The Roman mills were adopted throughout much of medieval Europe, and waterwheels of increasing size, made almost entirely of wood, were built until the 18th century.

Energy from ocean tides

In addition to flowing stream water, ocean tides were used to drive waterwheels. Tidal water was allowed to flow into large millponds, controlled initially through lock-type gates and later through flap valves. Once the tide ebbed, water was let out through sluice gates and directed onto the wheel. Sometimes the tidal flow was assisted by building a dam across the estuary of a small river. Although limited in operation to ebbing tide conditions, tidal mills were widely used by the 12th century. The earliest recorded reference to tidal mills is found in the *Domesday Book* (1086), which also records more than 5,000 water mills in England south of the Severn and Trent rivers. (Tidal mills also were built along the Atlantic coast in Europe and centuries later on the eastern seaboard of the United States and in Guyana, where they powered sugarcane-crushing mills.)

The first analysis of the performance of waterwheels was published in 1759 by John Smeaton, an English engineer. Smeaton built a test apparatus with a small wheel (its diameter was only 0.61 metre) to measure the effects of water velocity, as well as head and wheel speed. He found that the maximum efficiency (work produced divided by potential energy in the water) he could obtain was 22 percent for an undershot wheel and 63 percent for an overshot wheel (*i.e.*, one in which water enters the wheel above its centre; see Figure 1). In 1776 Smeaton became the first to use a cast-iron wheel, and two years later he introduced cast-iron gearing, thereby bringing to an end the all-wood construction that had prevailed since Roman times. Based on his model tests, Smeaton built an undershot wheel for the London Bridge waterworks that measured 4.6 metres wide and that had a diameter of 9.75 metres. The results of Smeaton's experimental work came to be widely used throughout Europe for designing new wheels.

Influence of Smeaton's work

During the mid-1700s a reaction waterwheel for generating small amounts of power became popular in the rural areas of England. In this type of device, commonly known as a Barker's mill, water flowed into a rotating vertical tube before being discharged through nozzles at the end of two horizontal arms. These directed the water out tangentially, much in the way that a modern rotary lawn sprinkler does. A rope or belt wound around the vertical tube provided the power takeoff.

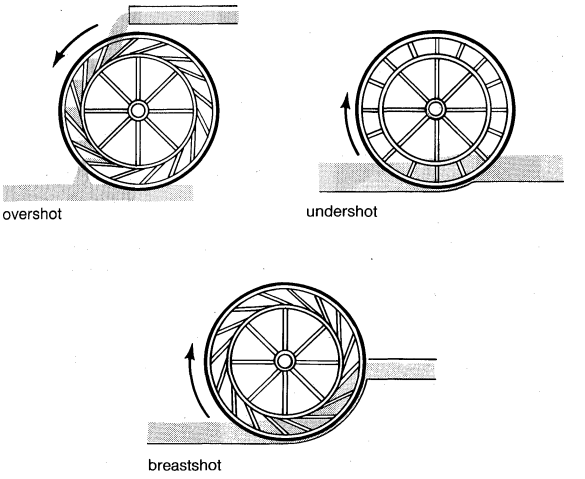


Figure 1: Some major types of waterwheels.

Early in the 19th century Jean-Victor Poncelet, a French mathematician and engineer, designed curved paddles for undershot wheels to allow the water to enter smoothly. His design was based on the idea that water would run up the surface of the curved vanes, come to rest at the inner diameter, and then fall away with practically no velocity. This design increased the efficiency of undershot wheels to 65 percent. At about the same time, William Fairbairn, a Scottish engineer, showed that breast wheels (*i.e.*, those in which water enters at the 10- or two-o'clock position) were more efficient than overshot wheels and less vulnerable to flood damage. He used curved buckets and provided a close-fitting masonry wall to keep the water from flowing out sideways. In 1828 Fairbairn introduced ventilated buckets in which gaps at the bottom of each bucket allowed trapped air to escape. Other improvements included a governor to control the sluice gates and spur gearing for the power takeoff.

During the course of the 19th century, waterwheels were slowly supplanted by water turbines (see *Water turbines* below). Water turbines were more efficient; design improvements eventually made it possible to regulate the speed of the turbines and to run them fast enough to drive electric generators. This fact notwithstanding, waterwheels gave way slowly, and it was not until the early 20th century that they became largely obsolescent. Yet, even today some waterwheels still survive; in the early 1970s there were more than 1,000 grain mills in use in Portugal alone. Equipped with submerged bearings, these modern waterwheels certainly are more sophisticated than their predecessors, though they bear a remarkable likeness to them.

Windmills. Windmills, like waterwheels, were among the original prime movers that replaced animal muscle as a source of power. They were used for centuries in various parts of the world, converting the energy of the wind into mechanical energy for grinding grain, pumping water, and draining lowland areas.

First known wind-powered device

The first known wind device was described by Hero of Alexandria (*c.* 1st century AD). It was modeled on a water-driven paddle wheel and was used to drive a piston pump that forced air through a wind organ to produce sound. The earliest known references to wind-driven grain mills, found in Arabic writings of the 9th century AD, refer to a Persian millwright of AD 644, although windmills may actually have been used earlier. These mills, erected near what is now the Iran–Afghanistan border, had a vertical shaft with paddlelike sails radiating outward and were located in a building with diametrically opposed openings

for the inlet and outlet of the wind. Each mill drove a single set of stones without gearing. The first mills were built with the millstones above the sails, patterned after the early waterwheels from which they were derived. Similar mills were known in China by the 13th century.

Windmills with vertical sails on horizontal shafts reached Europe through contact with the Arabs. Adopting the ideas from contemporary waterwheels, builders began to use fabric-covered, wood-framed sails located above the millstone, instead of a waterwheel below, to drive the grindstone through a set of gears. The whole mill with all its machinery was supported on a fixed post so that it could be rotated and faced into the wind. The millworks were initially covered by a boxlike wooden frame structure and later often by a "round-house," which also provided storage. A brake wheel on the shaft allowed the mill to be stopped by a rim brake. A heavy lever then had to be raised to release the brake, an early example of a fail-safe device. Mills of this sort first appeared in France in 1180, in areas of Syria under the control of the crusaders in 1190, and in England in 1191. The earliest known illustration is from the Windmill Psalter made in Canterbury, Eng., in the second half of the 13th century.

Emergence of the tower mill

The large effort required to turn a post-mill into the wind probably was responsible for the development of the so-called tower mill in France by the early 14th century (see Figure 2). Here, the millstone and the gearing were placed in a massive fixed tower, often circular in section and built of stone or brick. Only an upper cap, normally made of wood and bearing the sails on its shaft, had to be rotated. Such improved mills spread rapidly throughout Europe and later became popular with early American settlers.

The Low Countries of Europe, which had no suitable streams for waterpower, saw the greatest development of windmills. Dutch hollow post-mills, invented in the early 15th century, used a two-step gear drive for drainage pumps. An upright shaft that had gears on the top and bottom passed through the hollow post to drive a paddle-wheel-like scoop to raise water. The first wind-driven sawmill, built in 1592 in the Netherlands by Cornelis Cornelisz, was mounted on a raft to permit easy turning into the wind.

At first both post-mills and the caps of tower mills were turned manually into the wind. Later small posts were placed around the mill to allow winching of the mill with a chain. Eventually winches were placed into the caps of tower mills, engaged with geared racks and operated from inside or from the ground by a chain passing over a wheel. Tower mills had their sail-supporting or tail pole normally inclined at between 5° and 15° to the horizontal. This aided the distribution of the huge sail weight on the tail bearing and also provided greater clearance between the sails and the support structure. Windmills became progressively larger, with sails from about 17 to 24 metres in diameter already common in the 16th century. The material of construction, including all gearing, was wood, although eventually brass or gunmetal came into use for the main bearings. Cast-iron drives were first introduced in 1754 by John Smeaton, the aforementioned English engineer. Little is known about the actual power produced by these mills. In all likelihood only from 10 to 15 horsepower was developed at the grinding wheels. A 50-horsepower mill was not built until the 19th century. The maximum efficiency of large Dutch mills is estimated to have been about 20 percent.

Invention of the fantail

In 1745 Edmund Lee of England invented the fantail, a ring of five to eight vanes mounted behind the sails at right angles to them. These were connected by gears to wheels running on a track around the cap of the mill. As the wind changed direction, it struck the sides of the fantail vanes, realigning them and thereby turning the main sails again squarely into the wind. Fabric-on-wood-frame sails were sometimes replaced by all-wood sails with removable sections. Early sails had a constant angle of twist; variable twist sails resembling a modern airplane propeller were developed much later.

A major problem with all windmills was the need to feather the sails or reduce sail area so that if the wind suddenly increased during a storm the sails would not be ripped apart. In 1772 Andrew Meikle, a Scottish millwright, invented the spring sail, a shutter arrangement similar to a venetian blind in which the sails were controlled by a spring. When the wind pressure exceeded a preset amount, the shutters opened to let some of the wind pass through. In 1789 Stephen Hooper of England introduced roller blinds that could all be simultaneously adjusted with a manual chain from the ground while the mill was working. This was improved upon in 1807 by Sir William Cubitt, who combined Meikle's shutters with Hooper's remote control by hanging varying weights on the adjustment chain, thus making the control automatic. These so-called patent sails, however, found acceptance only in England and northern Europe.

Even though further improvements were made, especially in speed control, the importance of windmills as a major power producer began to decline after 1784, when the first flour mill in England successfully substituted a steam engine for wind power. Yet, the demise of windmills was slow; at one time in the 19th century there were as many as 900 corn (maize) and industrial windmills in the Zaan district of the Netherlands, the highest concentration known. Windmills persisted throughout the 19th century in newly settled or less-industrialized areas, such as the central and western United States, Canada, Australia, and New Zealand. They also were built by the hundreds in the West Indies to crush sugarcane.

Rex Wales

Figure 2: (Left) Post-mill with four "common sails," the cloths of which are fully set, at Marck, Pas-de-Calais, Fr. (Centre) Hollow post-mill with boarded sails, at Ylojärvi, Häme, Fin. (Right) Tower mill with patent sails and fantail, at Pakenham, Suffolk, Eng.

Use of wind pumps

The primary exception to the steady abandonment of windmills was resurgence in their use in rural areas for pumping water from wells. The first wind pump was introduced in the United States by David Hallay in 1854. After another American, Stewart Perry, began constructing wind pumps made of steel and equipped with metal vanes in 1883, this new and simple device spread around the world.

Wind-driven pumps remain important today in many rural parts of the world. They continued to be used in large numbers, even in the United States, well into the 20th century until low-cost electric power became readily available in rural areas. Although rather inefficient, they are rugged and reliable, need little attention, and remain a prime source for pumping small amounts of water wherever electricity is not economically available. (For the development of the modern wind turbine, see *Wind turbines* below.) (R.W./Fr.L.)

DEVELOPMENTS OF THE INDUSTRIAL REVOLUTION

Steam engines. The rapid growth of industry in Britain from about the mid-18th century (and somewhat later in various other countries) created a need for new sources of motive power, particularly those independent of geographic location and weather conditions. This situation, together with certain other factors, set the stage for the development and widespread use of the steam engine, the first practical device for converting thermal energy to mechanical energy.

Papin's experimental work on steam power

The foundations for the use of steam power are often traced to the experimental work of the French physicist Denis Papin. In 1679 Papin invented a type of pressure cooker, a closed vessel with a tightly fitting lid that confined steam until high pressure was generated. Observing that the steam in the vessel raised the lid, he conceived the idea of using steam to power a piston and cylinder engine.

Thomas Savery, an English inventor and military engineer, studied Papin's work and built a steam-driven suction machine for removing water from coal mines. Savery's machine (patented in 1698) consisted of a boiler, a closed, water-filled reservoir, and a series of valves. Steam was introduced into the reservoir, and the pressure of the steam forced the water out through a one-way outlet valve until the vessel was empty. Water was then sprayed over the surface of the vessel to condense the steam and create a vacuum capable of drawing up more water through a valve below. Unfortunately the vacuum created was not perfect, and so water could only be lifted to a limited height.

Newcomen engine. Some years later another English engineer, Thomas Newcomen, developed a more efficient steam pump consisting of a cylinder fitted with a piston—a design inspired by Papin's aforementioned idea. When the cylinder was filled with steam, a counterweighted pump plunger moved the piston to the extreme upper end of the stroke. With the admission of cooling water, the steam condensed, creating a vacuum. The atmospheric pressure in the mine acted on the piston and caused it to move down in the cylinder, and the pump plunger was lifted by the resulting force (see Figure 3).

Because Savery had obtained a broad patent for his steam device, Newcomen could not patent his engine. He thus entered into a partnership with Savery, and together they built, in 1712, the first piston-operated steam pump. Several years later Smeaton improved the Newcomen engine, almost doubling its efficiency. Although engines of this kind converted only about 1 percent of the thermal energy in the steam to mechanical energy, they remained unrivaled for more than 50 years.

First piston-operated steam pump

Watt's engine. In 1765 James Watt, a Scottish instrument maker and inventor, modified a Newcomen engine by adding a separate condenser to make it unnecessary to heat and cool the cylinder with each stroke. Because the cylinder and piston remained at steam temperature while the engine was operating, fuel costs dropped by about 75 percent.

Watt entered into a partnership with Matthew Boulton, who owned a factory in Soho, near Birmingham, Eng. At Boulton's insistence he set out to develop a new kind of engine that rotated a shaft instead of providing simple up-and-down motion. He found a way to obtain an inflexible connection between piston and rod (beam) and invented special gear arrangements to convert the up-and-down movement of the beam into circular motion. A heavy flywheel was added to smooth out the variations in the force delivered to the engine shaft by the action of the piston in the cylinder. The flow of steam to the engine was regulated by a governor connected to the flywheel. In addition, Watt applied steam to both sides of the piston to produce greater uniformity of effort and increased power.

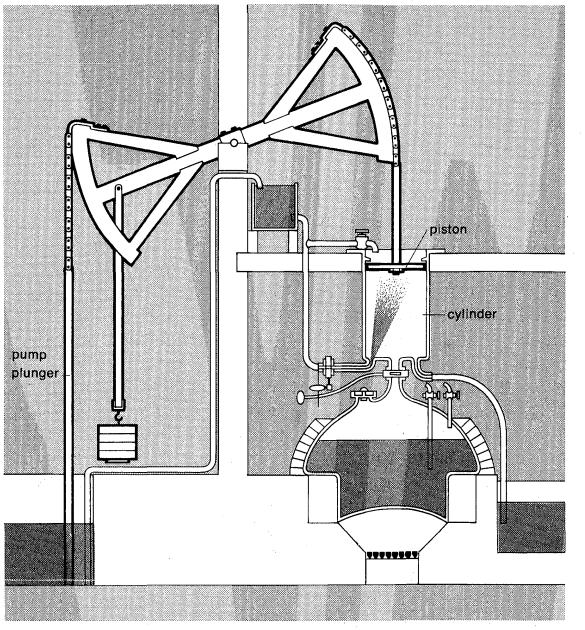


Figure 3: Newcomen engine (see text).

Widespread use of Watt's rotative steam engine

Although far more difficult to build, Watt's rotative engine opened up an entirely new field of application: it enabled the steam engine to be used to operate rotary machines in factories and cotton mills. The rotative engine was widely adopted; it is estimated that by 1800 Watt and Boulton had built 500 engines, of which less than 40 percent were pumps and the rest were of the rotative type.

High-pressure steam engines. Although Watt understood the advantages of utilizing the expansive power of steam within a cylinder, he refused to use steam under high pressure for reasons of safety. This limited the application of steam engines. By the early years of the 19th century, however, the American inventor Oliver Evans had built a stationary high-pressure steam engine for driving a rotary crusher to produce pulverized limestone for agricultural use. Within a few years Evans had designed lighter-weight high-pressure steam engines that could do various other tasks, such as drive sawmills, sow grain, and power a dredge. From 1806 to about 1816 he produced more than 100 steam engines that were employed with screw presses for processing paper, cotton, and tobacco.

Other major advances in the use of high-pressure steam were achieved by Richard Trevithick in England during the early years of the 19th century. Trevithick built the world's first steam-powered railway locomotive in 1803. Two years later he adapted his high-pressure steam engine to drive an iron-rolling mill and to propel a barge with the help of paddle wheels.

First steam locomotive engine

Watt's engine was able to convert only a little more than 2 percent of the thermal energy in steam to work. The improvements introduced by Evans, Trevithick, and others (*e.g.*, three separate expansion cycles and higher steam temperatures) increased the efficiency of the steam engine to roughly 17 percent by 1900. Yet, within the next decade the steam engine was supplanted for various important applications by the more efficient steam turbine (see *Steam turbines* below). Owing to technological advances and the use of high-temperature steam, steam turbines have attained an efficiency of thermal energy conversion of approximately 40 percent. (E.B.Wo./Ed.)

Stirling engine. Many of the early high-pressure steam

boilers exploded because of poor materials and faulty methods of construction. The resultant casualties and property losses motivated Robert Stirling of Scotland to invent a power cycle that operated without a high-pressure boiler. In his engine (patented in 1816), air was heated by external combustion through a heat exchanger and then was displaced, compressed, and expanded by two pistons. Stirling also conceived the idea of a regenerator to store thermal energy during part of the cycle and then return this energy to the working fluid. A successful Stirling engine was built for factory use in 1843, but general use was restricted by the high cost of the device. Nevertheless, until about 1920, small engines of this type were used to pump water on farms and to generate electricity for small communities.

External-combustion engine

Since the Stirling engine is efficient, produces less pollution than most other kinds of engines, and operates on virtually any kind of fuel, efforts have been made intermittently since the late 1930s to reduce its manufacturing costs. Modern versions of the Stirling engine employ pressurized hydrogen or helium instead of air. Although attempts were made as recently as the 1970s to adapt the device to power automobiles, its only commercial application at present is use as a cryogenic refrigerator.

Internal-combustion engines. While the steam engine remained dominant in industry and transportation during much of the 19th century, engineers and scientists began developing other sources and converters of energy. One of the most important of these was the internal-combustion engine. In such a device a fuel and oxidizer are burned within the engine and the products of combustion act directly on piston or rotor surfaces. By contrast, an external-combustion device, such as the steam engine, employs a secondary working fluid that is interposed between the combustion chamber and power-producing elements. By the early 1900s the internal-combustion engine had replaced the steam engine as the most broadly applied power-generating system not only because of its higher thermal efficiency (there is no transfer of heat from combustion gases to a secondary working fluid that results in losses in efficiency) but also because it provided a low-weight, reasonably compact, self-contained power plant.

First practical internal-combustion engine

The German engineer Nikolaus August Otto is generally credited with having built the first practical internal-combustion engine (1876), though several rudimentary devices had appeared earlier in the century. In 1885 Gottlieb Daimler, another German engineer, modified the four-cycle Otto engine so that it burned gasoline (instead of coal powder) and built the first successful high-speed internal-combustion engine. Within several decades the gasoline engine found wide application in motorcycles, automobiles, and small trucks (see *Gasoline engines* below).

Another type of internal-combustion engine was introduced by Rudolf Diesel, also of Germany, in the early 1890s. Named for its inventor, the diesel engine was more efficient than engines of the Otto variety and was fueled by heavy oil, which is cheaper and less volatile than gasoline. As a result, it was adopted as the primary power plant for submarines, railway locomotives, and heavy machinery (see *Diesel engines* below).

An internal-combustion engine quite different from the reciprocating piston type was developed around the turn of the century. This was the gas-turbine engine, the first successful version of which was built in 1903 in France. Modern gas turbines have been used for electric power generation and various other purposes, but its primary application has been jet propulsion. In a gas-turbine system compressed air, heated by the combustion of petroleum, is used to turn a turbine to drive the compressor while excess energy accelerates the exhaust gas to high velocity for producing thrust (see *Gas-turbine engines* and *Jet engines* below).

Early studies of rocket propulsion systems

Another form of propulsive engine, the rocket, attracted increasing attention during the final decades of the 19th century due in part to the imaginative portrayals of space travel fabricated by Jules Verne and other science-fiction writers. From about 1880, various scientists and inventors began investigating theoretical problems of rocket motion and propulsion system design. By the mid-1920s Robert H. Goddard of the United States had developed experimental rockets employing liquid and solid propellants (see *Rockets* below).

Electric generators and motors. Other important energy-conversion devices emerged during the 19th century. During the early 1830s the English physicist and chemist Michael Faraday discovered a means by which to convert mechanical energy into electricity on a large scale. While engaged in experimental work on magnetism, Faraday found that moving a permanent magnet into and out of a coil of wire induced an electric current in the wire. This process, called electromagnetic induction, provided the working principle for electric generators.

During the late 1860s Zénobe-Théophile Gramme, a French engineer and inventor, built a continuous-current generator. Dubbed the Gramme dynamo, this device contributed much to the general acceptance of electric power. By the early 1870s Gramme had developed several other dynamos, one of which was reversible and could be used as an electric motor. Electric motors, which convert electrical energy to mechanical energy, run virtually every kind of machine that uses electricity.

The Gramme dynamo

All of Gramme's machines were direct-current (DC) devices. It was not until 1888 that Nikola Tesla, a Serbian-American inventor, introduced the prototype of the present-day alternating-current (AC) motor (see *Electric generators and electric motors* below).

Direct energy-conversion devices. Most of these energy converters, sometimes called static energy-conversion devices, use electrons as their "working fluid" in place of the vapour or gas employed by such dynamic heat engines as the external-combustion and internal-combustion engines mentioned above. In recent years, direct energy-conversion devices have received much attention because of the necessity to develop more efficient ways of transforming available forms of primary energy into electric power. Four such devices—the electric battery, the fuel cell, the thermoelectric generator (or at least its working principle), and the solar cell—had their origins in the early 1800s.

Electrochemical generation of electricity

The battery, invented by the Italian physicist Alessandro Volta about 1800, changes chemical energy directly into an electric current. A device of this type has two electrodes, each of which is made of a different chemical. As chemical reactions occur, electrons are released on the negative electrode and made to flow through an external circuit to the positive electrode. The process continues until the circuit is interrupted or one of the reactants is exhausted. The forerunners of the modern dry cell and the lead-acid storage battery appeared during the second half of the 19th century (see *Batteries* below).

The fuel cell, another electrochemical producer of electricity, was developed by William Robert Grove, a British physicist, in 1839. In a fuel cell, continuous operation is achieved by feeding fuel (*e.g.*, hydrogen) and an oxidizer (oxygen) to the cell and removing the reaction products (see *Fuel cells* below).

Thermoelectricity

Thermoelectric generators are devices that convert heat directly into electricity. Electric current is generated when electrons are driven by thermal energy across a potential difference at the junction of two conductors made of dissimilar materials. This effect was discovered by Thomas Johann Seebeck, a German physicist, in 1821. Seebeck observed that a compass needle near a circuit made of different conducting materials was deflected when one of the junctions was heated. He investigated various materials that produce electric energy with an efficiency of 3 percent. This efficiency was comparable to that of the steam engines of the day. Yet, the significance of the discovery of the thermoelectric effect went unrecognized as a means of producing electricity because of Seebeck's misinterpretation of the phenomenon as a magnetic effect caused by a difference in temperature. A basic theory of thermoelectricity was finally formulated during the early 1900s, though no functional generators were developed until much later (see *Thermoelectric power generators* below).

Conversion of solar energy into electric power

In a solar cell, radiant energy drives electrons across a potential difference at a semiconductor junction in which the concentrations of impurities are different on the two sides of the junction. What is often considered the first

genuine solar cell was built in the late 1800s by Charles Fritts, who used junctions formed by coating selenium (a semiconductor) with an extremely thin layer of gold (see *Exploiting renewable energy sources* below).

MODERN DEVELOPMENTS

The 20th century brought a host of important scientific discoveries and technological advances, including new and better materials and improved methods of fabrication. These developments permitted the enhancement and refinement of many of the energy-conversion devices and systems that had been introduced during the previous century, as exemplified by the remarkable evolution of jet engines and rockets. They also gave rise to entirely new technologies.

Discovery and application of nuclear energy. *Fission reactors.* Scientists first learned of the tremendous energy bound in the nucleus of the atom during the early years of the century. In 1942 they succeeded in unleashing that energy on a large scale by means of what was called an atomic pile. This was the first nuclear fission reactor, a device designed to induce a self-sustaining and controlled series of fission reactions that split heavy nuclei to release their energy. It was built for the U.S. Manhattan Project undertaken to develop the atomic bomb. Shortly after World War II, reactors were built for submarine propulsion and for commercial power production. The first full-scale commercial nuclear power plant was opened in 1956 at Calder Hall, Eng. In a power generation system of this kind, much of the energy released by the fissioning of heavy nuclei (principally those of the radioactive isotope uranium-235) takes the form of heat, which is used to produce steam. This steam drives a turbine, the mechanical energy of which is converted to electricity by a generator (see *Nuclear fission reactors* below).

First nuclear reactor

Fusion reactors. In the late 1930s Hans A. Bethe, a German-born physicist, recognized that the fusion of hydrogen nuclei to form deuterium releases energy. Since that time scientists have sought to harness such thermonuclear reactions for practical energy production. Much of their work has centred on the use of magnetic fields and electromagnetic forces to confine plasma, an exceedingly hot gas composed of unbound electrons, ions, and neutral atoms and molecules. Plasma is the only state of matter in which thermonuclear reactions can be induced and sustained to generate usable amounts of thermal energy. The difficulty is in confining plasma long enough for this to happen. Although researchers have made significant headway toward constructing fusion reactors capable of such confinement, no device of this kind has been developed sufficiently for commercial application (see *Fusion reactors* below).

Other conversion technologies. Energy requirements for space vehicles led to an intensive investigation, from 1955 on, of all possible energy sources. Direct energy-conversion devices are of interest for providing electric power in spacecraft because of their reliability and their lack of moving parts. Besides solar cells, fuel cells, and thermoelectric generators, thermionic power converters have received considerable attention for space applications. Thermionic generators are designed to convert thermal energy directly into electricity. The required heat energy may be supplied by chemical, solar, or nuclear sources, the latter being the preferred choice for current experimental units (see *Thermionic power converters* below).

Energy-conversion research associated with the space program

Another direct energy converter with considerable potential is the magnetohydrodynamic (MHD) power generator. This system produces electricity directly from a high-temperature, high-pressure electrically conductive fluid—usually an ionized gas—moving through a strong magnetic field. The hot fluid may be derived from the combustion of coal or other fossil fuel. The first successful MHD generator was built and tested during the 1950s. Since that time developmental efforts have progressed steadily, culminating in a Soviet project to build an MHD power plant in the city of Ryazan, located about 180 kilometres (112 miles) southeast of Moscow (see *Magnetohydrodynamic power generators* below).

Exploiting renewable energy sources. Growing concern over the world's ever-increasing energy needs and the prospect of rapidly dwindling reserves of oil, natural gas, and uranium fuel have prompted efforts to develop viable alternative energy sources. The volatility and uncertainty of the petroleum fuel supply were dramatically brought to the fore during the energy crisis of the 1970s caused by the abrupt curtailment of oil shipments from the Middle East to many of the highly industrialized nations of the world. It also has been recognized that the heavy reliance on fossil fuels has had an adverse impact on the environment. Gasoline engines and steam-turbine power plants that burn coal or natural gas emit substantial amounts of sulfur dioxide and nitrogen oxides into the atmosphere. When these gases combine with atmospheric water vapour, they form sulfuric acid and nitric acids, giving rise to highly acidic precipitation. The combustion of fossil fuels also releases carbon dioxide. The amount of this gas in the atmosphere has steadily risen since the mid-1800s largely as a result of the growing consumption of coal, oil, and natural gas. More and more scientists believe that the atmospheric buildup of carbon dioxide (along with that of other industrial gases such as methane and chlorofluorocarbons) may induce a greenhouse effect, raising the surface temperature of the Earth by increasing the amount of heat trapped in the lower atmosphere. This condition could bring about climatic changes with serious repercussions for natural and agricultural ecosystems. (For a detailed discussion of acid rain and the greenhouse effect, see the articles ATMOSPHERE: *Effects of human activity on atmospheric composition and their ramifications* and HYDROSPHERE, THE: *Acid rain* and *Buildup of greenhouse gases.*)

Growing concern over depletion of fuel resources and environmental pollution

Many countries have initiated programs to develop renewable energy technologies that would enable them to reduce fossil-fuel consumption and its attendant problems. Fusion devices are believed to be the best long-term option, since their primary energy source would be the hydrogen isotope deuterium abundantly present in ordinary water. Other technologies that are being actively pursued are those designed to make wider and more efficient use of the energy in sunlight, wind, moving water, and terrestrial heat (*i.e.,* geothermal energy). The amount of energy in such renewable and virtually pollution-free sources is large in relation to world energy needs, yet at the present time only a small portion of it can be converted to electric power at reasonable cost.

A variety of devices and systems has been created to better tap the energy in sunlight. Among the most efficient are photovoltaic systems that transform radiant energy from the Sun directly into electricity by means of silicon or gallium arsenide solar cells. Large arrays consisting of thousands of these semiconductor cells can function as central power stations (see *Solar cells* below). Other systems, which are still under development, are designed to concentrate solar radiation not only to generate electric power but also to produce high-temperature process heat for various applications. These systems employ a number of different components, including large parabolic concentrators and heat engines of the Stirling engine type (see above). Another approach involves the use of flat-plate solar collectors to provide space heating for commercial and residential buildings.

Although wind is intermittent and diffuse, it contains tremendous amounts of energy. Sophisticated wind turbines have been developed to convert this energy to electric power. The utilization of wind energy systems grew discernibly during the 1980s. For example, more than 15,000 wind turbines are now in operation in Hawaii and California at specially selected sites. Their combined power rating of 1,500 megawatts is roughly equal to that of a conventional steam-turbine power installation (see *Wind turbines* below).

Renewal of interest in wind power

Converting the energy in moving water to electricity has been a long-standing technology. Yet, hydroelectric power plants are estimated to provide only about 2 percent of the world's energy requirements. The technology involved is simple enough: hydraulic turbines change the energy of fast-flowing or falling water into mechanical energy that drives power generators, which produce electricity

(see *Water turbines* below). Hydroelectric power plants, however, generally require the building of costly dams. Another factor that limits any significant increase in hydroelectric power production is the scarcity of suitable sites for additional installations except in certain regions of the world.

In certain coastal areas of the world, as, for example, the Rance River estuary in Brittany, Fr., hydraulic turbine-generator units have been used to harness the great amount of energy in ocean tides (see *Tidal plants* below). At most such sites, the capital costs of constructing damlike structures with which to trap and store water are prohibitive, however.

Geothermal energy flows from the hot interior of the Earth to the surface in steam or hot water most often in areas of active volcanism. Geothermal reservoirs with temperatures of 180° C or higher are suitable for power generation. The earliest commercial geothermal power plant was built in 1904 in Larderello, Italy. Today, steam from wells drilled to depths of hundreds of metres drives the plant's turbine generators to produce about 190 megawatts of electricity. Geothermal plants have been built in a number of other countries, including Japan, Mexico, New Zealand, the Soviet Union, and the United States. The principal U.S. plant, located at The Geysers north of San Francisco, can generate up to 1,900 megawatts, though production may be restricted to prolong the life of the steam field. (C.R.R./Ed.)

First commercial geothermal power plant

MAJOR ENERGY-CONVERSION DEVICES AND SYSTEMS

Turbines

A turbine is a machine that converts the energy stored in a fluid into mechanical energy. This conversion is generally accomplished by passing the fluid through a system of stationary passages or vanes that alternate with passages consisting of finlike blades attached to a rotor. By arranging the flow so that a tangential force, or torque, is exerted on the rotor blades, the rotor will turn, and work can be extracted. Turbines can be classified into four general types according to the fluids used: water, steam, gas, and wind. Although the same principles apply to all turbines, their specific designs differ sufficiently to merit separate descriptions.

Classification of turbines

A water turbine uses the potential energy resulting from the difference in elevation between an upstream water reservoir and the turbine-exit water level (the tailrace) to convert this so-called head into work. Water turbines are the modern successors of simple waterwheels, which date back about 2,000 years (see *Waterwheels* above). Today, the primary use of water turbines is for electric power generation.

The greatest amount of electrical energy comes, however, from steam turbines coupled to electric generators. The turbines are driven by steam produced in either a fossil-fuel-fired or a nuclear-powered generator. The energy that can be extracted from the steam is conveniently expressed in terms of the enthalpy change across the turbine. Enthalpy reflects both thermal and mechanical energy forms in a flow process and is given by the sum of the internal thermal energy and the product of pressure times volume. The available enthalpy change through a steam turbine increases with the temperature and pressure of the steam generator and with reduced turbine-exit pressure.

Enthalpy

For gas turbines, the energy extracted from the fluid also can be expressed in terms of the enthalpy change, which for a gas is nearly proportional to the temperature drop across the turbine. In gas turbines the working fluid is air mixed with the gaseous products of combustion. Most gas-turbine engines include at least a compressor, a combustion chamber, and a turbine. These are usually mounted as an integral unit and operate as a complete prime mover on a so-called open cycle where air is drawn in from the atmosphere and the products of combustion are finally discharged again to the atmosphere. Since successful operation depends on the integration of all components, it is important to consider the whole device, which is actually an internal-combustion engine, rather than the turbine alone. For this reason, gas turbines will be treated in the section *Internal-combustion engines.*

The energy available in wind can be extracted by a wind turbine to produce electric power or to pump water from wells. Wind turbines are the successors of windmills, which were important sources of power from the late Middle Ages through the 19th century (see *Windmills* above). (Fr.L.)

WATER TURBINES

Principal types

Water turbines are generally divided into two categories: (1) impulse turbines used for high heads of water and low flow rates and (2) reaction turbines normally employed for heads below about 450 metres and moderate or high flow rates. These two classes include the main types in common use—namely, the Pelton impulse turbine and the reaction turbines of the Francis, propeller, Kaplan, and Deriaz variety. Turbines can be arranged with either horizontal or, more commonly, vertical shafts. Wide design variations are possible within each type to meet the specific local hydraulic conditions. Today, most hydraulic turbines are used for generating electricity in hydroelectric installations.

Impulse turbines. In an impulse turbine the potential energy, or the head of water, is first converted into kinetic energy by discharging water through a carefully shaped nozzle. The jet, discharged into air, is directed onto curved buckets fixed on the periphery of the runner to extract the water energy and convert it to useful work.

Modern impulse turbines are based on a design patented in 1889 by the American engineer Lester Allen Pelton. The free water jet strikes the turbine buckets tangentially. Each bucket has a high centre ridge so that the flow is divided to leave the runner at both sides. Pelton wheels are suitable for high heads, typically above about 450 metres with relatively low water flow rates. For maximum efficiency the runner tip speed should equal about one-half the striking jet velocity. The efficiency (work produced by the turbine divided by the kinetic energy of the free jet) can exceed 91 percent when operating at 60–80 percent of full load.

Pelton turbines

The power of a given wheel can be increased by using more than one jet. Two-jet arrangements are common for horizontal shafts (see Figure 4). Sometimes two separate runners are mounted on one shaft driving a single electric generator. Vertical-shaft units may have four or more separate jets.

If the electric load on the turbine changes, its power output must be rapidly adjusted to match the demand. This

Fox Photos Ltd.

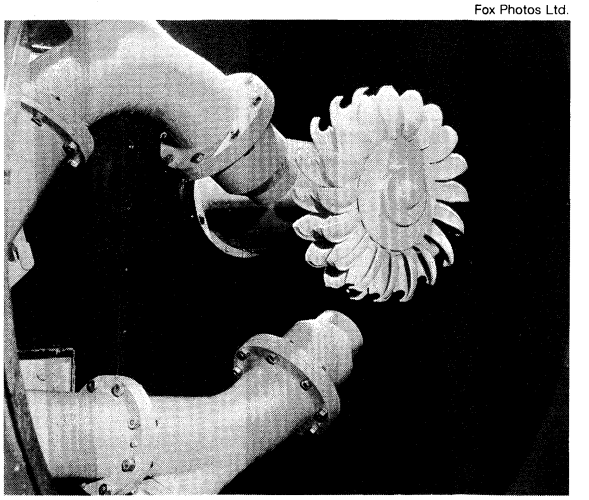


Figure 4: Pelton water turbine with twin jets.

requires a change in the water flow rate to keep the generator speed constant. The flow rate through each nozzle is controlled by a centrally located, carefully shaped spear or needle that slides forward or backward as controlled by a hydraulic servomotor.

Proper needle design assures that the velocity of the water leaving the nozzle remains essentially the same irrespective of the opening, assuring nearly constant efficiencies over much of the operating range. It is not prudent to reduce the water flow suddenly to match a load decrease. This could lead to a destructive pressure surge (water hammer) in the supply pipeline, or penstock. Such surges can be avoided by adding a temporary spill nozzle that opens while the main nozzle closes or, more commonly, by partially inserting a deflector plate between the jet and the wheel, diverting and dissipating some of the energy while the needle is slowly closed.

Turgo impulse turbines

Another type of impulse turbine is the turgo type. The jet impinges at an oblique angle on the runner from one side and continues in a single path, discharging at the other side of the runner. This type of turbine has been used in medium-sized units for moderately high heads.

Reaction turbines. In a reaction turbine, forces driving the rotor are achieved by the reaction of an accelerating water flow in the runner while the pressure drops. The reaction principle can be observed in a rotary lawn sprinkler where the emerging jet drives the rotor in the opposite direction. Due to the great variety of possible runner designs, reaction turbines can be used over a much larger range of heads and flow rates than impulse turbines. Reaction turbines typically have a spiral inlet casing that includes control gates to regulate the water flow. In the inlet a fraction of the potential energy of the water may be converted to kinetic energy as the flow accelerates. The water energy is subsequently extracted in the rotor.

There are, as noted above, four major kinds of reaction turbines in wide use: the Kaplan, Francis, Deriaz, and propeller type. In fixed-blade propeller and adjustable-blade Kaplan turbines (named after the Austrian inventor Victor Kaplan), there is essentially an axial flow through the machine. The Francis- and Deriaz-type turbines (after the British-born American inventor James B. Francis and the Swiss engineer Paul Deriaz, respectively) use a "mixed flow," where the water enters radially inward and discharges axially. Runner blades on Francis and propeller turbines consist of fixed blading, while in Kaplan and Deriaz turbines the blades can be rotated about their axis, which is at right angles to the main shaft.

Propeller-type turbines

Axial-flow machines. Fixed propeller-type turbines are generally used for large units at low heads, resulting in large diameters and slow rotational speeds. As the name suggests, a propeller-type turbine runner looks like the very large propeller of a ship except that it serves the opposite purpose: power is extracted in a turbine, whereas it is fed into a marine propeller. The central shaft, or hub, may have the propeller blades bolted to it during on-site assembly, thus permitting shipment by sections for a large runner. At low heads (below about 24 metres), vertical-shaft propeller turbines typically have a concrete spiral inlet casing of rectangular cross section. Inlet guide vanes are either mounted on a ring or, in large units, set individually directly into the concrete. The flow passage can be increased or decreased by servomotor-driven wicket gates. The kinetic energy leaving the runner can be partially recaptured by a draft tube, a conical diffusing exit section where the velocity is decreased while the pressure is increased. This leads to improved efficiency by keeping the loss of kinetic energy in the exit, or tail, section of the installation to a minimum.

Propeller turbines are used extensively in North America, where low heads and large flow rates are common. For example, there are 32 propeller turbines in the Moses–Saunders Power Dam on the St. Lawrence River between New York and Ontario—16 operated by the United States and 16 by Canada, with each turbine rated at 50,000 kilowatts. With such large plants it is possible to run each turbine at or near its most efficient output by switching complete units in or out as the load fluctuates, in addition to regulating each unit.

If the head or the water flow rate tends to vary seasonally, as occurs in many river systems, an installation with only a few propeller turbines might have to operate all units at partial output under average flow and load conditions. The energy-conversion efficiency of a conventional propeller turbine decreases rapidly once the turbine load drops below 75 percent of its rating. This performance loss can be minimized by varying the inlet-blade angle of the runner to match the runner-inlet conditions more accurately with the water velocity for a given flow. In such a Kaplan turbine (Figure 5) each blade can be swiveled about a post at right angles to the main turbine shaft, thus producing a variable pitch. The angle of the blades is controlled by an oil-pressure operated servomotor, usually mounted in the rotor hub with the oil fed through the generator and turbine shaft. The servo-control system, which also drives the gates through a cam or rocker arrangement, is designed to adjust angles and inlet flows to match the electrical load while keeping the main shaft with its directly coupled generator rotating at constant speed. Runners with four to six blades are common, though more blades may be used for high heads. British manufacturers have developed Kaplan designs for heads up to 58 metres.

Kaplan turbines

By courtesy of The English Electric Co. Ltd.

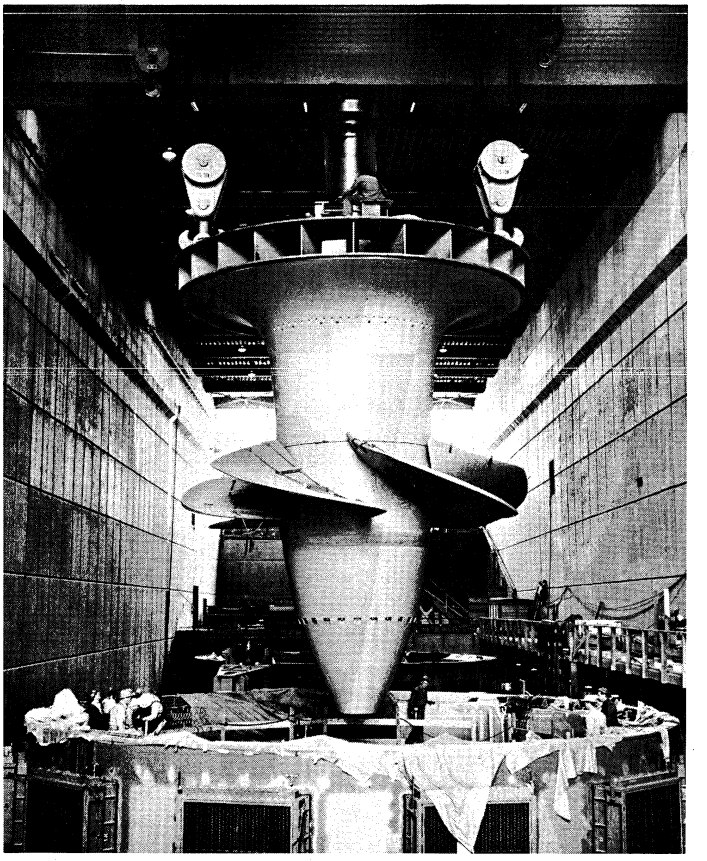


Figure 5: A 131,000-horsepower Kaplan water-turbine runner.

Although the usual turbine installation has a vertical shaft, some also have been designed with horizontal shafts. In a horizontal bulb arrangement, the generator is embedded in a nacelle, corresponding to the thick body of a light bulb, while the blades are set around a hub corresponding to the thinner bulb socket. This design is suitable for medium-sized machines operating at very low heads when an almost straight-through water flow is desirable. The Rance River tidal plant in France employs this kind of arrangement (see *Tidal plants* below).

Francis turbines

Mixed-flow turbines. Francis turbines are probably used most extensively because of their wider range of suitable heads, characteristically from three to 600 metres. At the high-head range, the flow rate and the output must be large; otherwise the runner becomes too small for reasonable fabrication. At the low-head end, propeller turbines are usually more efficient unless the power output is also

small. Francis turbines reign supreme in the medium-head range of 120 to 300 metres and come in a wide range of designs and sizes. They can have either horizontal or vertical shafts, the latter being used for machines with diameters of about two metres or more. Vertical-shaft machines usually occupy less space than horizontal units, permit greater submergence of the runner with a minimum of deep excavation, and make the tip-mounted generator more easily accessible for maintenance. Horizontal-shaft units are more compact for smaller sizes and allow easier access to the turbine, although removal of the generator for repair becomes more difficult as size increases.

The most common form of Francis turbine has a welded, or cast-steel, spiral casing. The casing distributes water evenly to all inlet gates; up to 24 pivoted gates or guide vanes have been used. The gates operate from fully closed to wide open, depending on the power output desired. Most are driven by a common regulating speed ring and are pin-connected in such a fashion that no damage will occur if debris blocks one of the gate passages. The regulating ring is rotated by one or two oil-pressure servomotors that are controlled by the speed governor.

Slow, high-power units have a nearly radial set of blades, while in fast and lower-powered units the curved blades reach from the radial inlet to almost the axial outlet (see Figure 6). Once the overall blade dimensions (inlet and exit diameters and blade height) have been defined, the blades are designed for a smooth entry of the water flow at the inlet and minimum water swirl at the exit. The number of blades can vary from seven to 19. Runners for low-head units are usually made of cast mild steel, sometimes with stainless-steel protection added at locations subject to cavitation (see below). All stainless-steel construction is more commonly used for high heads. Large units can be welded together on-site, using an appropriate combination of various preformed steel sections to provide carefully shaped, finished water passages. Francis turbines allow for very large, high-output units. The Grand Coulee hydroelectric power plant on the Columbia River in Washington state has the largest single runner in the United States, a device capable of producing 716,000 kilowatts at a head of 93 metres. The world's largest hydroelectric power installation, the Itaipú plant on the Paraná River between Brazil and Paraguay, is scheduled to have 18 Francis turbines capable of producing 740,000 kilowatts each at heads between 118.4 and 126.7 metres while rotating at slightly above 90 revolutions per minute (rpm).

By courtesy of Voith Hydro, Inc.

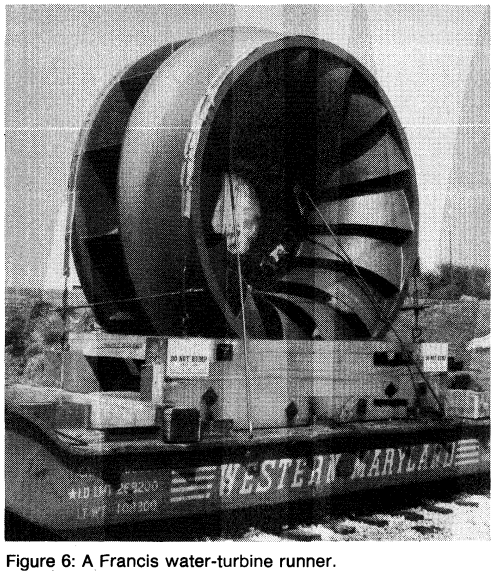


Figure 6: A Francis water-turbine runner.

Deriaz-type mixed-flow turbine

A mixed-flow turbine of the Deriaz type uses swiveled, variable-pitch runner blades that allow for improved efficiency at part loads in medium-sized machines (see Figure 7). The Deriaz design has proved useful for higher heads and also for some pumped storage applications (see below). It has the advantage of a lower runaway (sudden loss of load) speed than a Kaplan turbine, which results in significant savings in generator costs. Very few Deriaz turbines, however, have actually been built. The first non-reversible Deriaz turbine, capable of producing 22,750 kilowatts with a head of 55 metres, was installed in an underground station at Culligran, Scot., in 1958.

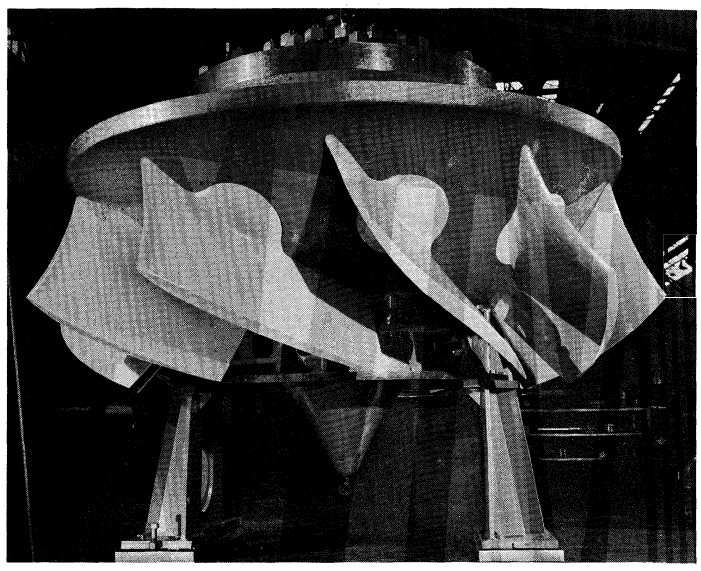


Figure 7: Runner for 108,000-horsepower Deriaz reversible pump turbine for Valdecañas Power Station, Spain.
By courtesy of The English Electric Co. Ltd.

Other design considerations. *Output and speed control.* If the load on the generator is decreased, a turbine will tend to speed up unless the flow rate can be reduced accordingly. Similarly, an increase of load will cause the turbine to slow down unless more water can be admitted. Since electric-generator speeds must be kept constant to a high degree of precision, this leads to complex controls. These must take into account the large masses and inertias of the metal and the flowing water, including the water in the inflow pipes (or penstocks), that will be affected by any change in the wicket gate setting. If the inlet pipeline is long, the closing time of the wicket gate must be slow enough to keep the pressure increase caused by a reduction in flow velocity within acceptable limits. If the closing or opening rate is too slow, control instabilities may result. To assist regulation with long pipelines, a surge chamber is often connected to the pipeline as close to the turbine as possible. This enables part of the water in the line to pass into the surge chamber when the wicket gates are rapidly closed or opened. Medium-sized reaction turbines may also be provided with pressure-relief valves through which some water can be bypassed automatically as the governor starts to close the turbine. In some applications, both relief valves and surge chambers have been used.

Cavitation. According to Bernoulli's principle (derived by the Swiss mathematician Daniel Bernoulli), as the flow velocity of the water increases at any given elevation, the pressure will drop. There is a danger that in high-velocity sections of a reaction turbine, especially near the exit, the pressure can become so low that the water flashes over into small vapour bubbles, which then collapse suddenly. This so-called cavitation leads to erosion pitting as well as to vibrations and must be avoided by the careful shaping of all blade passages and of the exit passage or draft tube.

Problems associated with cavitation

Turbine selection on the basis of specific speed. Initial turbine selection is usually based on the ratio of design variables known as the power specific speed. In U.S. design practice this is given by

$$N = \frac{nP^{1/2}}{H^{5/4}},$$

where n is in revolutions per minute, P is the output in horsepower, and H is the head of water in feet. Turbine types can be classified by their specific speed, N, which always applies at the point of maximum efficiency. If N ranges from one to 20, corresponding to high heads and

low rotational speeds, impulse turbines are appropriate. For N between 10 and 90, Francis-type runners should be selected, with slow-running, near-radial units for the lower N values and more rapidly rotating mixed-flow runners for higher N values. For N up to 110, Deriaz turbines may be suitable. If N ranges from 70 to the maximum of 260, propeller or Kaplan turbines are called for.

Using the specific speed formula, a turbine designed to deliver 100,000 horsepower (74,600 kilowatts) with a head of 40 feet (12.2 metres) operating at 72 revolutions per minute would have a specific speed of 226, suggesting a propeller or Kaplan turbine. It can also be shown that the flow rate would have to be about 24,500 cubic feet per second (694 cubic metres per second) at a turbine efficiency of 90 percent. The runner diameter will be about 33 feet (10 metres). This illustrates the large sizes required for high-power, low-head installations and the low rotational speed at which these turbines have to operate to stay within the permissible specific speed range.

Turbine model testing. Before building large-scale installations, the design should be checked out with turbine model tests, using geometrically similar models of small and intermediate size, all operating at the same specific speed. Allowances must be made for the effects of friction, determined by the Reynolds number (density × rotational speed × runner diameter squared/viscosity) and for possible changes in scaled roughness and clearance dimensions. Friction effects are less important for large units, which tend to be more efficient than smaller ones.

Applications. *Electric power generation.* Water turbines are used almost exclusively for generating electric power that can be transmitted through high-voltage power lines to population centres. The United States, the Soviet Union, and Canada lead in hydroelectric power production, though many other countries also have major production facilities. Until the late 1950s most single turbogenerator units had capacities of less than 150,000 kilowatts. By the late 1980s construction costs and the need for reliability pointed toward 250,000- to 300,000-kilowatt units, although some recent installations were equipped with turbines capable of up to 750,000 kilowatts.

Pumped storage. Electricity must be used as soon as it is generated; there are no economical means of storing large quantities of electric energy. Thus hydroelectric plants built for near-maximum power consumption during daytime peak hours would have to operate at low efficiency during nighttime or weekend off-hours. To avoid this, water can be pumped to a second, higher reservoir during off-hours for storage in the form of potential energy and then fed back through power-generating turbines at times of high demand. Even though this system does not generate new energy (there actually is a reduction in energy due to losses involved in pump and turbine operation as well as in the electric motor and generator), pumped hydro-storage often becomes economical when compared with the cost of constructing additional turbines for peak power demands.

Use of reversible-pump turbines

Modern pumped storage units in the United States normally use reversible-pump turbines that can be run in one direction as pumps and in the other direction as turbines. These are coupled to reversible electric motor/generators. The motor drives the pump during the storage portion of the cycle, while the generator produces electricity during discharge from the upper reservoir.

Most reversible-pump turbines are of the Francis type. The complexity of the unit, however, increases significantly as compared to a turbine alone. In spite of the higher costs for both hydraulic and electrical controls and support equipment, the total installed cost will be less than for completely separate pump-motor and turbine-generator assemblies with dual water passages.

Some very economical pumped storage plants have heads exceeding 300 metres. In the past this was considered too high for single-stage pumps, and the use of separate multistage, nonreversible units was required. Satisfactory reversible single-stage pump turbines, however, have been developed that can operate at 700-metre heads, though most installations have smaller head differences between the upper and lower reservoirs.

For medium heads, Deriaz turbines have had some success because they allow ready adjustment of the runner-blade angles to match the opposite requirements of pumping and power generation. The pumping load can also be varied with Deriaz-type units, which cannot be done with a Francis runner. A further advantage of a Deriaz-type machine is that the runner blades can be closed to form a smooth cone, a feature that permits pump start-up with minimum load while the unit is submerged in water.

An early major Deriaz reversible-pump turbine system was installed at plants on both the Canadian and U.S. sides of Niagara Falls; this made it possible to provide "side storage" at night without impairing the tourist attraction of the falls by reducing the flow during the day. The Tuscarora plant on the U.S. side uses 12 pump turbines at heads between 18.3 to 29 metres.

Pumped storage has become widespread in industrialized nations. In the United States alone more than 30 pumped hydropower stations were in operation by the mid-1980s. The largest plant is located in Bath County, Va., where six pump-turbines have a total capacity of 2.1 million kilowatts. This amount of power can be generated over an 11-hour period.

Tidal plants. Although the majority of hydroelectric plants depend on the impoundment of rivers, tidal power still could play a role, albeit minor, in electric power generation during the coming years. Areas where the normal tide runs high, such as in the Bay of Fundy between the United States and Canada or along the English Channel, can allow water to flow into a dam-controlled basin during high tide and discharge it during low tide to produce intermittent power. One such plant is located in France on the estuary of the Rance River near Saint-Malo in Brittany. There, a reservoir has been created by a barrage four kilometres inland from the river mouth to make use of tides ranging from about 3.4 to 13.4 metres. The power station is equipped with 24 reversible bulb-type propeller turbines coupled to reversible motor/generators, each having a capacity of 10,000 kilowatts. Pumped storage is used if the tidal outflow through the plant falls below peak power demands. A pilot tidal plant with a 40,000-kilowatt capacity has been built in the Soviet Union on the Barents Sea. If this facility proves economical, it may lead to the construction of other tidal plants on the northern and eastern Soviet coasts.

Rance River tidal plant

Cost of hydroelectric power. Although large hydroelectric plants can be operated economically, the cost of land acquisition and of dam and reservoir construction must be included in the total cost of power, since these outlays generally account for about half of the total initial cost. Most large plants serve multiple purposes: hydropower generation, flood control, storage of drinking water, and the impounding of water for irrigation. By properly prorating construction costs to the non-power-producing utility of the unit, electricity can be sold very cheaply. In the Pacific Northwest region of the United States, such accounting has given hydroelectric plants an apparent cost advantage over fossil-fueled units.

History of water turbine technology. Experiments on the mechanics of reaction wheels conducted by the Swiss mathematician Leonhard Euler and his son Albert in the 1750s found application about 75 years later. In 1826 Jean-Victor Poncelet of France proposed the idea of an inward-flowing radial turbine, the direct precursor of the modern water turbine. This machine had a vertical spindle and a runner with curved blades that was fully enclosed. Water entered radially inward and discharged downward below the spindle.

Precursor of the modern water turbine

A similar machine was patented in 1838 by Samuel B. Howd of the United States and built subsequently. Howd's design was improved on by James B. Francis, who added stationary guide vanes and shaped the blades so that water could enter shock-free at the correct angle. His runner design, which came to be known as the Francis turbine (see above), is still the most widely used for medium-high heads. Improved control was proposed by James Thomson, a Scottish engineer, who added coupled and pivoted curved guide vanes to assure proper flow directions even at part load.

A radial outward-flow turbine had been proposed in 1824 by the French engineering professor Claude Burdin and his former student Benoît Fourneyron. This device had a vertical axis carrying a runner with curved blades through which the water left almost tangentially. Fixed guide vanes, curved in the opposite direction, were mounted in an annulus inside the runner. Unfortunately the design made it difficult to support the runner and to take power off the turbine wheel. The first successful version of the turbine was built by Fourneyron in 1827. More than 100 such machines were subsequently built all over the world; they achieved efficiencies up to 75 percent at full load with heads up to 107 metres. In 1844 Uriah A. Boyden added an outlet diffuser to recover part of the kinetic energy exiting the device and thereby further improved efficiency. Outward-flow turbines, however, are inherently unstable, and speed control is difficult. Moreover, the construction of outward-flow turbines is very complex as compared to that of Francis-type runners, and this fact led to their eventually being supplanted by the latter.

Francis turbines were augmented by the development of the Pelton wheel (1889) for small flow rates and high heads and by propeller turbines, first built by Kaplan in 1913, for large flows at low heads. Kaplan's variable-pitch propeller turbine, which still bears his name, was manufactured after 1920. These units, together with the Deriaz mixed-flow turbine (invented in 1956), constitute the arsenal of modern water turbines.

By the mid-19th century, water turbines were widely used to drive sawmills and textile mill equipment, often through a complex system of gears, shafts, and pulleys. After the widespread adoption of the steam engine they did not, however, become a major factor in power generation until the advent of the electric generator made hydroelectric power possible.

The world's first hydroelectric central station

The world's first hydroelectric central station was built in 1882 in Appleton, Wis., only three years after Thomas Edison's invention of the light bulb. Its output of 12.5 kilowatts was used to light two paper mills and a house. Thereafter hydroelectric power development spread rapidly, though even by 1910 most units delivered only a few hundred to a few thousand kilowatts. Installations with more than 100,000-kilowatt capacity were not built until the 1930s. One of the first large U.S. plants was installed at Hoover Dam on the Colorado River between Nevada and Arizona. It began operating in 1936 and eventually included 17 Francis turbines capable of delivering from 40,000 to 130,000 kilowatts of power, along with two 3,000-kilowatt Pelton wheels.

Pumped-storage hydropower in the United States

The first pumped storage plant with a capacity of 1,500 kilowatts was built near Schaffhausen, Switz., in 1909. It made use of a separate pump and turbine, resulting in a relatively large and only barely economical system. The first U.S. plant, built on the Rocky River in Connecticut in 1929, was also only marginally economical. In the United States major work on pumped-storage hydropower began in the mid-1950s, following the success of a plant at Flatiron, Colo. Built in 1954, this facility was equipped with a reversible-pump turbine having a capacity of 9,000 kilowatts.

In highly industrialized countries, such as the United States and the nations of western Europe, most potential sites for hydropower have already been tapped. Environmental concerns relating to the impact of large dams on the upstream watercourse and to the possible effect on aquatic life add to the likelihood that only a few large hydraulic plants will be built in the future.

From about the 1940s to the early 1970s, many small U.S. hydroelectric facilities (primarily those of less than 1,000-kilowatt capacity) were, in fact, closed down because high maintenance and supervision costs made them uneconomical compared to power plants that burn fossil fuels. Even though the increase in fossil-fuel costs since 1973 has led to the rehabilitation of some of these abandoned plants, only a marked increase in fuel prices, coupled with specific needs for irrigation or flood control, is likely to lead to significant new hydroelectric plant construction.

It is estimated that about 75 percent of the potential waterpower in the contiguous United States has already been developed, with the drainage area of the Columbia River in the Pacific Northwest leading in both developed and potential additional power. As of the late 1980s, hydroelectric power met about 13 percent of the total demand for electrical energy in the United States, though this amounts to only 3 percent of the combined U.S. energy usage for mechanical power, heat, light, and refrigeration.

The above considerations do not necessarily apply to such remote areas as Alaska, northern Canada, and parts of the Soviet Union, or to developing nations in regions of the Himalayas, Africa, and South America. In these areas it is estimated that only 23 percent of the potential waterpower has been developed. For example, less than 1 percent of the estimated 167 million kilowatts available in Alaska has been harnessed to date. Other river basins with large remaining potential capacities include the Fraser River in Canada, the Orinoco in Venezuela, the Brahmaputra in India, and the Yenisey–Angara in the Soviet Union. Turbine capacities for some of these remote areas may possibly exceed the current maximum of 740,000 kilowatts per unit.

STEAM TURBINES

A steam turbine consists of a rotor resting on bearings and enclosed in a cylindrical casing. The rotor is turned by steam impinging against attached vanes or blades on which it exerts a force in the tangential direction. Thus a steam turbine could be viewed as a complex series of windmill-like arrangements, all assembled on the same shaft.

Predominance in power generation and propulsion of high-speed vessels

Because of its ability to develop tremendous power within a comparatively small space, the steam turbine has superseded all other prime movers, except hydraulic turbines, for generating large amounts of electricity and for providing propulsive power for large, high-speed ships. Today, units capable of generating more than 1.3 million kilowatts of power can be mounted on a single shaft.

Classifications. Large steam turbines are complex machines that can be classified in various ways. One approach centres on whether rotation is achieved by impulse forces or by reaction forces (see below). This distinction may become somewhat blurred, since many modern machines employ a combination of both methods.

Condensing and noncondensing turbines. Steam turbines are often divided into two types: condensing and noncondensing. In devices of the first type, steam is condensed at below atmospheric pressure so as to gain the maximum amount of energy from it. In noncondensing turbines, steam leaves the turbine at above atmospheric pressure and is then used for heating or for other required processes before being returned as water to the boiler. Compared to the fuel needed for simply converting water into steam (saturated steam), relatively little additional fuel has to be expended to increase the steam generator exit pressure and, especially, the temperature in order to produce superheated steam, which then is used to drive a turbine. Noncondensing turbines are thus an economical means of generating power (cogeneration) when large amounts of heating or process steam are already needed.

In condensing turbines, substantial quantities of cooling water are required to carry away the heat released during condensation. While noncondensing turbines exhaust steam at or above atmospheric pressure, condensing turbines can condense at pressures of 90 to 100 kilopascals (13 to 14.5 pounds per square inch) below atmospheric pressure. This allows for a much larger expansion of the steam and a larger change in enthalpy (see above), resulting in higher work output and greater efficiency. All central station plants, where efficiency is a prime consideration, employ condensing turbines.

Steam extraction. Steam turbines differ according to whether or not a portion of the steam is extracted from intermediate portions of the turbine. Extraction may be carried out to partially reheat the water fed back to the boiler and thereby significantly increase the efficiency of the power plant. In light of this, turbines may be classified as (1) straight-through turbines, in which there is no extraction (or bleeding), (2) bleeder or extraction turbines, and (3) controlled-(or automatic-) extraction turbines.

Increasing power plant efficiency

In bleeder turbines no effort is made to control the pres-

sure of the extracted steam, which varies in almost direct proportion to the load carried by the turbine. Extraction also reduces the steam flow to the condenser, allowing the turbine exhaust area to be reduced. Controlled-extraction turbines are designed for withdrawing variable amounts of constant-pressure steam irrespective of the load on the turbine. They are frequently selected for industrial use when steam at fixed intermediate pressures is demanded by process operations. Since both extraction pressures and turbine speed should be kept constant, a complex system is required for controlling steam flow, which increases the cost. Controlled-extraction turbines may be designed for both condensing and noncondensing operations.

Reheat and nonreheat turbines. If high-pressure, high-temperature steam is partially expanded through a turbine, the efficiency can be increased by returning the steam to the steam generator and reheating it to approximately its original temperature before feeding it back to the turbine. Single reheat turbines are common in the electric utility industry. For very large units, double reheating may be employed. Nonreheat turbines are currently limited mostly to industrial plants and small utilities.

Multiflow and compound arrangements. Steam entering a turbine at a high pressure and temperature—say, 24,100 kilopascals gauge, or 3,500 pounds per square inch gauge (where gauge denotes pressure above atmospheric value), and 600° C—can have a volume increase of more than a thousandfold if it is expanded to below atmospheric condenser pressures. To keep the steam velocity through the turbine essentially constant, the annular flow area would have to increase more than a thousandfold, necessitating very large diameter casings and excessively long turbine blades near the exit. In large turbines this problem is alleviated by splitting the low-pressure stream into a number of parallel flow sections, as illustrated in Figures 8 and 9 for four-flow units.

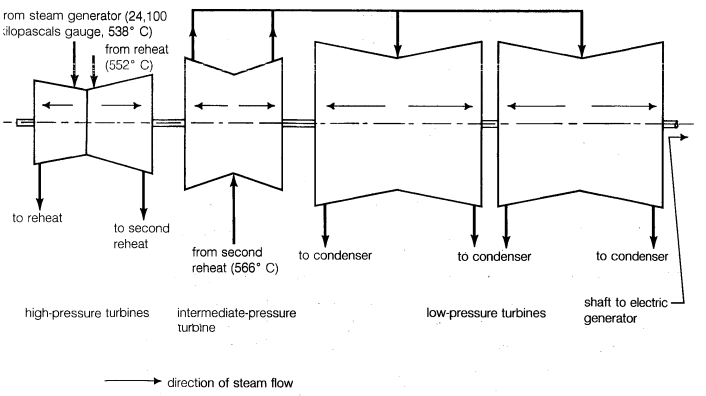


Figure 8: *Schematic of a 3,600-rpm, tandem-compound, four-flow, double-reheat steam turbine.* This unit can produce 725,000 kilowatts with inlet steam at 24,100 kilopascals gauge at 538° C and double reheats to 552° C and 566° C, respectively. The exit blades (or buckets) are 85 centimetres long. The heat rate (see text) at 100 percent of rated output is 7,490 British thermal units per kilowatt-hour (Btu/kWh), 7,950 Btu/kWh at 50 percent output, and 9,330 Btu/kWh at 25 percent output. The first reheat occurs at about 7,140 kilopascals gauge and the second at approximately 2,340 kilopascals gauge when operating at full power.

Tandem-compound and cross-compound turbines

This flow splitting also leads to another method of classification that differentiates between having the whole machine assembled along a single shaft with one generator (tandem-compound turbines), as illustrated in the figures, or utilizing two shafts, each with its own generator (cross-compound turbines).

Principal components. The main parts of a steam turbine are (1) the rotor that carries the blading to convert the thermal energy of the steam into the rotary motion of the shaft, (2) the casing, inside of which the rotor turns, that serves as a pressure vessel for containing the steam (it also accommodates fixed nozzle passages or stator vanes through which the steam is accelerated before being directed against and through the rotor blading), (3) the speed-regulating mechanism, and (4) the support system, which includes the lubrication system for the bearings that support the rotor and also absorb any end thrust developed.

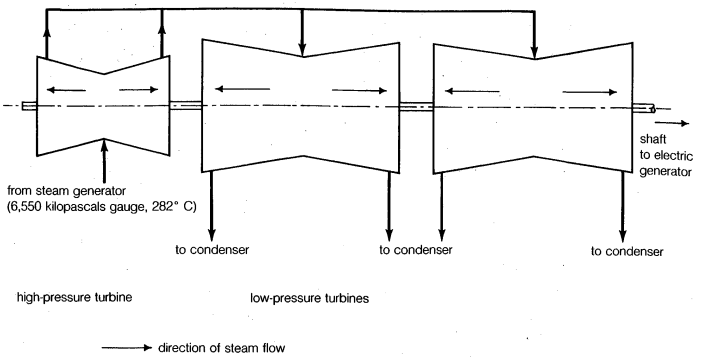


Figure 9: *Schematic of a 1,800-rpm, tandem-compound, four-flow, nonreheat steam turbine for nuclear applications.* The unit is rated at 847,000 kilowatts with steam entering at 6,550 kilopascals gauge and a temperature of about 282° C, which is typical for nuclear power plants. The exit buckets are 96.5 centimetres long. The heat rate is 9,810 Btu/kWh at full output and about 10,400 Btu/kWh at 40 percent output. Similarly, six-flow units operating with steam at 7,580 kilopascals gauge and about 293° C can produce 1,325,000 kilowatts. In this case the exit blades measure 109 centimetres long.

Design considerations. *Blading design.* The turbine blading must be carefully designed with the correct aerodynamic shape to properly turn the flowing steam and generate rotational energy efficiently. The blades also have to be strong enough to withstand high centrifugal stresses and must be sized to avoid dangerous vibrations. Various types of blading arrangements have been proposed, but all are designed to take advantage of the principle that when a given mass of steam suddenly changes its velocity, a force is then exerted by the mass in direct proportion to the rate of change of velocity.

Impulse and reaction blading

Two types of blading have been developed to a high degree of perfection: impulse blading and reaction blading. The principle of impulse blading is illustrated in the schematic diagram of Figure 10 for a first stage. A series of stationary nozzles allows the steam to expand to a lower pressure while its velocity and kinetic energy increase. The steam is then directed to the moving passages or buckets where the kinetic energy is extracted. Since there is ideally no pressure drop and no acceleration in the blade passage, the magnitude of the velocity vector in the blades should remain constant. This also implies that the cross-sectional area normal to the flow remains constant, giving rise to the typical shape of a symmetrical impulse blade—namely, thick at the middle and sharp at the ends.

Figure 10 also includes the velocity diagrams for such a stage. Velocities are vectors that are added by the parallelogram law (see ANALYSIS: *Vector and tensor analysis*). The relative velocity of the fluid with reference to the blade at inlet (or exit) added vectorially to the (tangential) velocity of the blade must give the absolute velocity as seen by the stationary passages. That the kinetic energy at the nozzle exit (proportional to the square of the nozzle-leaving velocity) is much larger than that at the blade exit is apparent from the figure. In an ideal impulse stage, this change of kinetic energy is fully converted into useful work. For minimum exit kinetic energy in a symmetrical impulse blade, the rotor velocity should be about one-half of the entering steam velocity.

In an idealized reaction stage, about one-half of the enthalpy drop per stage is effected in the stator passage and the other half in the rotor passage. This implies that the pressure drop is also almost equal in both the stationary and the rotary passages, which tend to look like mirror images of each other. If the flow velocity is subsonic (below the velocity of sound in the fluid), an expanding passage flow will increase its velocity as the pressure drops while the cross-sectional area decreases simultaneously, thus leading to the curved nozzle shape shown in Figure 11.

Since there is no pressure drop in an idealized impulse stage, pressure forces on the rotor play no role in this type of

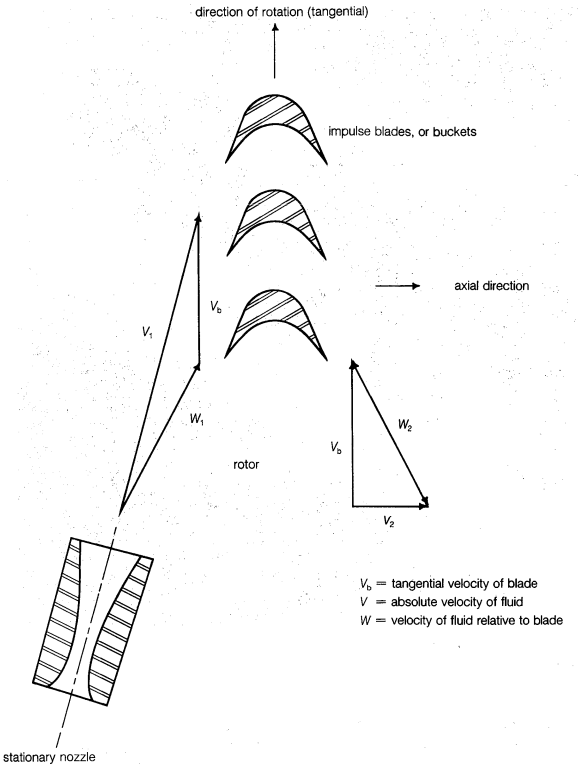


Figure 10: *Schematic of an impulse stage with velocity diagrams.*
The first stage, including a convergent-divergent inlet nozzle, is shown. Ideally there is no change in the magnitude of the relative velocities W between inlet and exit (which are designated by subscripts 1 and 2, respectively). The large inlet absolute velocity V_1 has been reduced to a small absolute exit velocity V_2, which ideally is in the axial direction.

arrangement. By contrast, in a reaction stage, the effect of the changing pressure exerts a net force in the tangential direction (thus turning the wheel) and also in the axial direction. The latter tends to push the rotor into the ends of the casing, requiring a thrust bearing to absorb the axial load. In large turbines the axial load can be reduced by admitting the steam flow in the middle and expanding in both axial directions, as shown in Figures 8 and 9.

There is no need to match the increase of fluid velocity in the stator to that in the rotor (50 percent reaction). Other

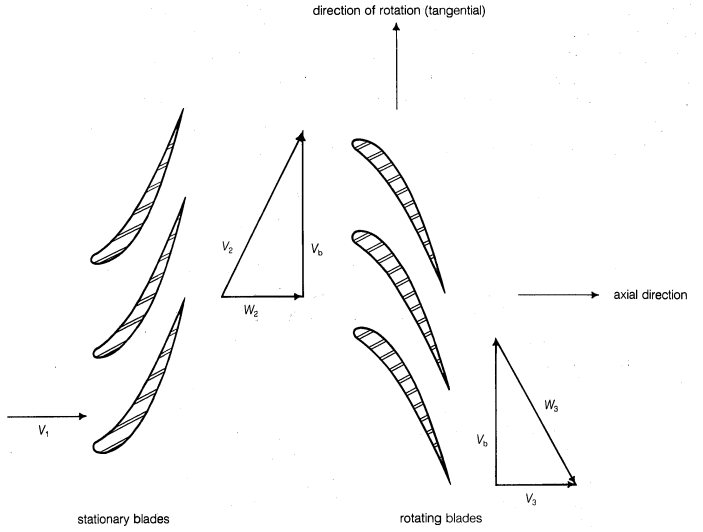


Figure 11: *An idealized 50-percent reaction stage for a steam turbine with velocity diagrams.*
Here, V is absolute velocity of fluid, V_b is blade velocity, and W is velocity of fluid relative to blade. Subscript 1 signifies entering stationary blade (stator), subscript 2 indicates leaving stator or entering rotor, and subscript 3 signifies leaving rotor.

widely used combinations that fall between pure impulse and 50 percent reaction staging have been developed.

The large length of low-pressure blades imposes special requirements on stiffness in addition to aerodynamic shaping. The tangential velocity of the blade near the hub is much smaller than at the blade tip, while the axial through-flow velocity is maintained nearly constant. To match the flow, the blades must be twisted to have the correct approach angle for the incoming steam (see Figure 12) and at the same time avoid possible resonant vibrations.

Turbine staging. Only a small fraction of the overall pressure drop available in a turbine can be extracted in a single stage consisting of a set of stationary nozzles or vanes and moving blades or buckets. In contrast to water turbines where the total head is extracted in a single runner (see above), the steam velocities obtained from the enthalpy drop between steam generator and condenser would be prohibitively high. In addition, the volume increase of the expanding steam requires a large increase in the annular flow area to keep the axial through-flow velocity nearly constant. To this must be added limitations on blade length and blade-tip velocities to avoid excessive centrifugal stresses. In practice, the steam expansion is therefore broken up into many small segments or stages, each with a range of velocities and an appropriate blade size to permit efficient conversion of the thermal energy in the steam to mechanical energy. In modern turbines, three types of staging are employed, either separately or in combination: (1) pressure (or impulse) staging, (2) reaction staging, and (3) velocity-compound staging.

By courtesy of GE Power Generation Operations, General Electric Co.

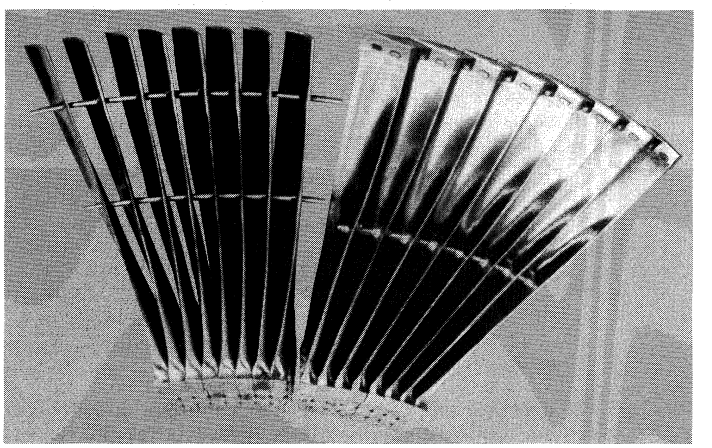


Figure 12: Large low-pressure steam turbine blades.

Pressure staging

Pressure staging uses a number of sequential impulse stages similar to those illustrated in Figure 10, except that the stationary passages also become highly curved nozzles. Pressure-staged turbines can range in power capacity from a few to more than 1.3 million kilowatts. Some manufacturers prefer to build units with impulse stages simply to reduce thrust-bearing loads. An example of a large turbine using impulse staging is shown in Figure 13. Such units may have as many as 20 sequential stages.

Reaction staging

Reaction staging is similar to pressure staging, except that a greater number of reaction stages are required. The first turbine stage, however, is often an impulse stage for controlling the steam flow and for rapidly reducing the pressure in stationary nozzles from its high steam generator value, thereby lowering the pressure that the casing has to withstand. Reaction turbines require about twice as many stages as impulse-staged turbines for the same change in steam enthalpy. The cost and size of the turbines, however, are about the same because blading for pressure staging must withstand greater forces and must therefore be more rigidly constructed. Reaction turbines also have large axial thrust and require heavy-duty thrust bearings.

Velocity-compound staging

In velocity-compound staging a set of stationary nozzles is followed by two sets of moving blades with a stationary row of impulse blades between them to redirect the flow. Ideally this allows twice as much power to be extracted than from a single impulse stage for a given blade-tip

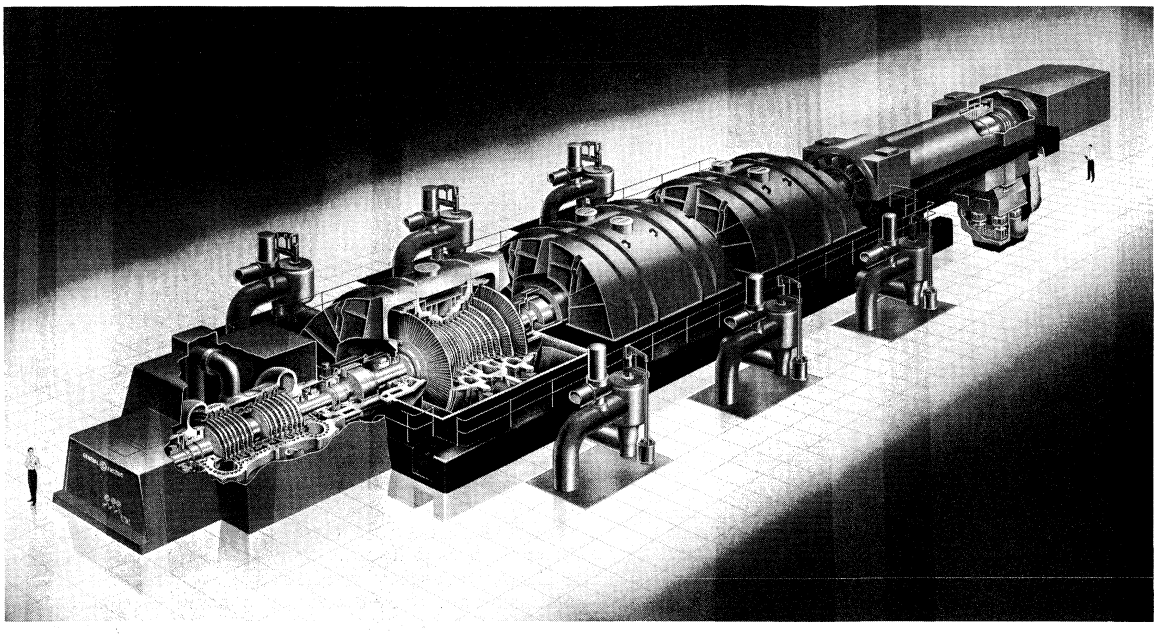


Figure 13: Rendering of an installed 1,800-revolution-per-minute tandem-compound, six-flow, nuclear steam turbine-generator unit rated at more than one million kilowatts.
By courtesy of General Electric Company, Schenectady, N.Y.

velocity. It also permits a large pressure drop through the stationary nozzles. Velocity-compounding is well suited for small turbines; it is also sometimes used as the first stage in large turbines for control purposes. The inherent high steam velocities, however, tend to result in high losses and poor stage efficiencies.

Power development. The theoretical maximum power produced by a turbine can be computed from the mass flow rate of the steam multiplied by the ideal enthalpy drop per unit mass between the steam generator exit and the condenser conditions. The actual power produced, however, is less because of friction, turbulence, leakage around the blade tips, and other losses. For the same maximum blade-tip velocity, pressure staging produces about twice as much ideal power per stage as reaction staging, while velocity-compound staging produces about four times as much.

The stage efficiency—*i.e.,* the amount of work that is actually produced in each stage as compared to the maximum possible amount—can be higher for reaction stages than for impulse stages due to generally lower flow velocities and associated losses. The greater number of stages required, however, results in an overall turbine efficiency that is about the same for both. Efficient stages also require carefully designed seals along the rotor shaft and opposite the rotating blade tips to avoid leakage past the blades.

Control. A turbine driving an electric generator must run at constant speed. In the United States where 60-cycle-per-second alternating current is used, this usually means 3,600 or 1,800 revolutions per minute. (In countries that use 50-cycle current, 3,000 or 1,500 revolutions per minute are the norm.) When the electric power demand on the generator, or the load, changes, the turbine must respond immediately to keep the speed constant. The inlet enthalpy is determined by the exit conditions of the steam generator and the exit enthalpy by the condenser pressure. Neither of these can be varied rapidly. With a fixed enthalpy drop per unit mass, the power output thus can only be controlled by varying the mass flow rate. This is achieved by opening or closing valves leading to the turbine inlet stage. Under partial load, the reduced steam flow results in lower axial velocities along the turbine and thereby alters the velocity diagrams somewhat. Since efficient operation requires a careful match between all velocity directions and blade inlet shapes, part-load operation decreases the efficiency of the turbine.

Overall performance characteristics. The performance of a steam turbine is conventionally measured in terms of its heat rate—*i.e.,* the amount of heat that has to be supplied to the feedwater in order to produce a specified generator power output. In the United States the heat rate is given by the heat input in Btus per hour for each kilowatt-hour of electricity produced by the turbogenerator assembly. The heat rate depends on the steam generator exit temperature and pressure, the condenser pressure, the efficiency of the turbine in converting the thermal energy of the steam into work, the mechanical and bearing losses, the exhaust loss due to the kinetic energy of the steam leaving the final turbine stage, and the generator losses. The lower the heat rate, the less the thermal energy required and the better the efficiency. At constant condenser pressure, the heat rate can be decreased by about 11 percent when going from steam generator exit conditions of 10,000 kilopascals gauge and 538° C to 24,100 kilopascals gauge and 538° C, with a subsequent reheat temperature of 538° C. The higher pressure, however, necessitates costlier equipment to contain the steam and to maintain the same reliability. Part-load operation, with its attendant loss of efficiency, always leads to higher heat rates.

Heat rate

History of steam turbine technology. *Early precursors.* The first device that can be classified as a reaction steam turbine is the aeolipile proposed by Hero of Alexandria, during the 1st century AD. In this device, steam was supplied through a hollow rotating shaft to a hollow rotating sphere. It then emerged through two opposing curved tubes, just as water issues from a rotating lawn sprinkler. The device was little more than a toy, since no useful work was produced.

Another steam-driven machine, described in 1629 in Italy, was designed in such a way that a jet of steam impinged on blades extending from a wheel and caused it to rotate by the impulse principle. Starting with a 1784 patent by James Watt, the developer of the steam engine, a number of reaction and impulse turbines were proposed, all adaptations of similar devices that operated with water. None were successful except for the units built by William Avery of the United States after 1837. In one such Avery turbine two hollow arms, about 75 centimetres long, were attached at right angles to a hollow shaft through which steam was supplied. Nozzles at the outer end of the arms allowed the steam to escape in a tangential direction, thus producing the reaction to turn the wheel. About 50 of these turbines were built for sawmills, cotton gins, and woodworking shops, and at least one was tried on a locomotive. While the efficiencies matched those of contemporary steam engines, high noise levels, difficult speed regulation, and frequent need for repairs led to their abandonment.

Avery turbine

Development of modern steam turbines. No further developments occurred until the end of the 19th century when various inventors laid the groundwork for the modern steam turbine. In 1884 Sir Charles Algernon Parsons, a British engineer, recognized the advantage of employing a large number of stages in series, allowing extraction of the thermal energy in the steam in small steps. Parsons also developed the reaction-stage principle according to which a nearly equal pressure drop and energy release takes place in both the stationary and moving blade passages. In addition, he subsequently built the first practical large marine steam turbines. During the 1880s Carl G.P. de Laval of Sweden constructed small reaction turbines that turned at about 40,000 revolutions per minute to drive cream separators. Their high speed, however, made them unsuitable for other commercial applications. De Laval then turned his attention to single-stage impulse turbines that used convergent-divergent nozzles, such as the one in Figure 14. From 1889 to 1897 de Laval built many turbines with capacities from about 15 to several hundred horsepower. His 15-horsepower turbines were the first employed for marine propulsion (1892). C.E.A. Rateau of France first developed multistage impulse turbines during the 1890s. At about the same time, Charles G. Curtis of the United States developed the velocity-compounded impulse stage.

By 1900 the largest steam turbine-generator unit produced 1,200 kilowatts, and 10 years later the capacity of such machines had increased to more than 30,000 kilowatts. This far exceeded the output of even the largest steam engines, making steam turbines the principal prime movers in central power stations after the first decade of the 20th century. Following the successful installation of a series of 68,000-horsepower turbines in the transatlantic passenger liners *Lusitania* and *Mauretania,* launched in 1906, steam turbines also gained preeminence in large-scale marine applications, first with vessels burning fossil fuels and then with those using nuclear power. Steam generator pressures increased from about 1,000 kilopascals gauge in 1895 to 1,380 kilopascals gauge by 1919 and then to 9,300 kilopascals gauge by 1940. Steam temperatures climbed from about 180° C (saturated steam) to 315° C (superheated steam) and eventually to 510° C over the same time period, while heat rates decreased from about 38,000 to below 10,000 Btus per kilowatt-hour.

The steam turbine as the principal prime mover in large power stations

Recent developments and trends. By 1940, single turbine units with a power capacity of 100,000 kilowatts were common. Ever-larger turbines (with higher efficiencies) have been constructed during the last half of the century, largely because of the steadily rising cost of fossil fuels. This required a substantial increase in steam generator pressures and temperatures. Some units operating with supercritical steam at pressures as high as 34,500 kilopascals gauge and at temperatures of up to 650° C were built before 1970. Reheat turbines that operate at lower pressures (between 17,100 to 24,100 kilopascals gauge) and temperatures (540–565° C) are now commonly installed to assure high reliability. Steam turbines in nuclear power plants, which are still being constructed in a number of countries outside of the United States, typically operate at about 7,580 kilopascals gauge and at temperatures of up to 295° C to accommodate the limitations of reactors. Turbines that exceed one-million-kilowatt output require exceptionally large, highly alloyed steel blades at the low pressure end.

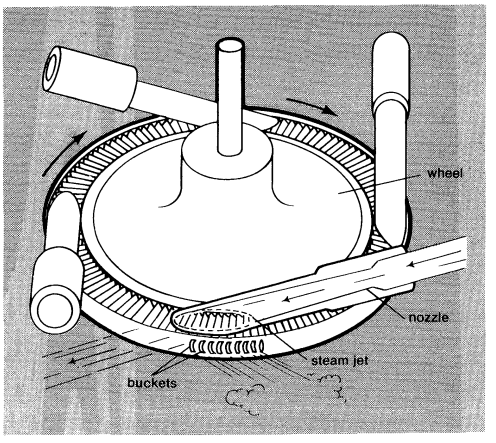


Figure 14: De Laval turbine, showing how the steam is formed into a jet by a specially shaped nozzle and is then deflected by the buckets or vanes on the wheel, causing the wheel to rotate.

Slightly more efficient units with a power capacity of more than 1.3 million kilowatts may eventually be built, but no major improvements are expected within the next few decades, primarily because of the temperature limitations of the materials employed in steam generators, piping, and high-pressure turbine components and because of the need for very high reliability.

Although the use of large steam turbines is tied to electric power production and marine propulsion, smaller units may be used for cogeneration when steam is required for other purposes, such as for chemical processing, powering other machines (*e.g.,* compressors of large central air-conditioning systems serving many buildings), or driving large pumps and fans in power stations or refineries. However, the need for a complete steam plant, including steam generators, pumps, and accessories, does not make the steam turbine an attractive power device for small installations. (R.A.B./Fr.L.)

WIND TURBINES

Modern wind turbines extract energy from the wind, mostly for electricity generation, by rotation of a propeller-like set of blades that drive a generator through appropriate shafts and gears. The older term windmill is often still used to describe this type of device, although electric power generation rather than milling has become the primary application. As was noted earlier, windmills, together with waterwheels, were widely used from the Middle Ages to the 19th century during the course of which they were supplanted by steam engines and steam turbines. Though they continued to be used for pumping water in rural areas, wind turbines practically disappeared in the 20th century as the internal-combustion engine and electricity provided more reliable and usually less expensive power. Interest in wind turbines for electricity generation was rekindled by the oil crisis of the mid-1970s. High initial costs, intermittent operation, and maintenance costs, however, have prevented wind turbines from becoming a significant factor in commercial power production.

Modern version of the windmill

Types of wind turbines. *Horizontal axis machines.* The best-known machines of this type are the so-called American farm windmills that came into wide use during the 1890s. Such devices consist of a rotor, which may have up to 20 essentially flat sheet-metal blades and a tail vane that keeps the rotor facing into the wind by swiveling the entire rotor assembly. Governing is automatic and overspeeding is avoided by turning the wheel off the wind direction, thus reducing the effective sail area while keeping the speed constant. A typical pump can deliver about 38 litres (10 gallons) per minute to a height of 30 metres at a wind velocity of 6.7 metres per second (15 miles per hour).

Modern wind turbines have from one to four metal blades that operate at much higher rotor-tip speeds than windmills. Each blade is twisted like an airplane propeller. An automatic governor rotates the blades about their support axis to maintain constant generator speed. The Jacobs three-bladed windmill, used widely between 1930 and 1960, could deliver about one kilowatt of power at a wind speed of 6.25 metres per second, a typical average wind velocity in the United States about 18 metres above ground.

High rotor-tip speeds

More recently, large horizontal-shaft, two-bladed turbines have been developed in the United States. The first such device, a unit equipped with a rotor measuring 11.6 metres in diameter, was installed near Sandusky, Ohio, in 1976; its power output was rated at 100 kilowatts. The most recent type of machine, first installed on the island of Oahu in Hawaii, has a rotor diameter of 122 metres with its axis about 76 metres above ground. Its output rating is 6,200 kilowatts at a wind speed of 13 metres per second.

Vertical-axis machines. Devices of this kind, which had

Savonius rotor

not been used since the early Middle Ages, found a new application after the Finnish engineer S.J. Savonius invented a new type of rotor in 1922. Known as the Savonius rotor, it consists of semicircular blades that can be constructed from little more than the two sections of an oil drum, cut in half along its vertical axis, and welded together with an offset from the axis to form an open S (see Figure 15). An advanced version of this machine installed at Manhattan, Kan., during the 1970s generated five kilowatts of electric power in a 12-metre-per-second wind.

By courtesy of Gary Johnson, Ph.D.

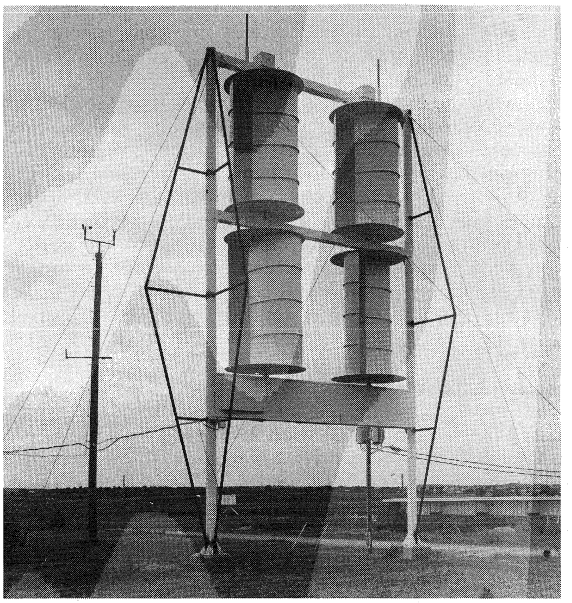


Figure 15: A Savonius rotor.

The most recent vertical wind turbine is based on a machine patented in 1931 by the French engineer G.J.M. Darrieus. Its two blades consist of twisted metal strips tied to the shaft at the top and bottom and bowed out in the middle similar to the blades on a food mixer. A Darrieus turbine with aluminum blades erected in 1980 by the Sandia National Laboratories in New Mexico produced 60 kilowatts in a wind blowing slightly more than 120 metres per second. Turbines of this variety are not self-starting and require an external motor for start-up. Several models of Darrieus turbines have been built since the construction of the Sandia unit (see Figure 16).

Wind farms. A wind farm is a cluster of wind turbines (up to several hundred) erected in areas where there is a nearly steady prevalent wind; such areas generally occur near mountain passes. Wind farms comprised of propeller-type units have been set up in Hawaii, California, and New Hampshire (see Figure 17). Capacities range from 10 to 500 kilowatts per unit. During 1984 the total output of all U.S. wind farms exceeded 150 million kilowatt-hours; the entire output was fed into the electric utility network. Though seemingly substantial, this amounted to less than $^{1}/_{100,000}$ of the total electric power generated in the United States.

Limitations on wind power. Not all the kinetic energy of the wind can be extracted because there must be a finite velocity as the air leaves the blading. It can be shown that the maximum efficiency (energy extracted divided by energy available in the captured wind area) obtainable is about 59 percent, although actual wind turbines extract only a portion of this amount. Currently, the maximum efficiency obtainable with a propeller-type windmill is roughly 47 percent; this occurs when the propeller-tip speed is between five and six times the wind velocity. For a given rotor speed, it drops rapidly as the wind velocity decreases. The power obtainable varies as the square of the rotor diameter and the cube of the wind velocity. Thus the theoretical maximum energy obtainable from a rotor with a diameter of 30 metres in a wind with a speed of 14 metres per second would be about 690 kilowatts. If the wind speed decreases to 7 metres per second, the theoretical maximum drops to about 86 kilowatts. At this lower wind speed, it would require more than 17,000 wind turbines (with rotors of 30 metres across) operating at an efficiency of 40 percent to match the output of a single large one-million-kilowatt central power station. When these limitations are coupled to the need for suitable sites with steady winds, it becomes apparent that wind turbines alone will not play a major role in meeting the power demands of an industrialized nation.

Amount of kinetic energy extracted from the wind

Development of wind turbines. The origin and development of the traditional windmill and other predecessors of modern wind turbines were described above in *History of energy-conversion technology.* The emergence and evolution of wind-driven devices for electric power generation are briefly surveyed here.

The development of the electric generator aroused some interest in the wind as a "free" power source. The first windmill to drive a generator was built in 1890 by P. LaCour in Denmark, using patent sails and twin fantails on a steel tower.

Adopting the ideas gained from airfoil and aircraft propeller designs, windmill designers and manufacturers began to replace broad windmill sails with a few slender propeller-like blades. In 1931 the first propeller wind turbine was erected in the Crimea. From the 1940s, experimental twin-blade turbines were constructed in the United States and later in Scotland and France. In The Netherlands a few old-fashioned mills were adapted to generate electricity. Today, wind turbines for electric power generation are most commonly propeller-type machines. (Fr.L.)

By courtesy of the U.S. Department of Energy

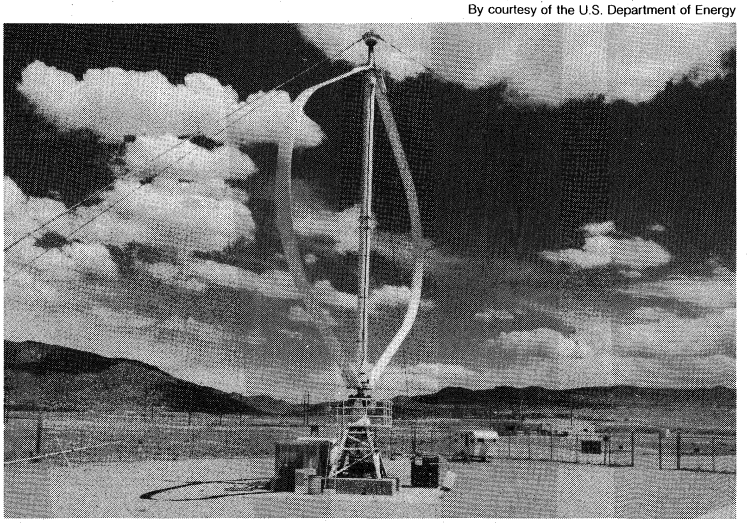


Figure 16: A Darrieus wind turbine.

Figure 17: A wind farm consisting of hundreds of propeller-type wind turbines.

Internal-combustion engines

An internal-combustion (IC) engine is any of a group of devices in which the reactants of combustion (oxidizer and fuel) and the products of combustion serve as the working fluids of the engine. Such an engine gains its energy from heat released during the combustion of the nonreacted working fluids, the oxidizer–fuel mixture. This process occurs within the engine and is part of the thermodynamic cycle of the device. Useful work generated by an IC engine results from the hot, gaseous products of combustion acting on moving surfaces of the engine, such as the face of a piston, a turbine blade, or a nozzle.

Major classes of IC engines

Internal-combustion engines are divided into two groups: continuous-combustion engines and intermittent-combustion engines. The continuous-combustion engine is characterized by a steady flow of fuel and oxidizer into the engine. A stable flame is maintained within the engine (*e.g.,* jet engine). The intermittent-combustion engine is characterized by periodic ignition of air and fuel and is commonly referred to as a reciprocating engine. Discrete volumes of air and fuel are processed in a cyclic manner. Gasoline piston engines and diesel engines are examples of this second group.

Internal-combustion engines can be delineated in terms of a series of thermodynamic events. In the continuous-combustion engine, the thermodynamic events occur simultaneously as the oxidizer and fuel, and the products of combustion flow steadily through the engine. In the intermittent-combustion engine, by contrast, the events occur in succession and are repeated for each full cycle.

With the exception of rockets (both solid-rocket motors and liquid-propellant rocket engines), internal-combustion engines ingest air, then either compress the air and introduce fuel into the air or introduce fuel and compress the air–fuel mixture, burn the air–fuel mixture, extract work from the hot, gaseous products of combustion by expansion, and ultimately exhaust the products of combustion. Their operation can be contrasted with that of external-combustion engines (*e.g.,* steam engines), in which the working fluid does not chemically react and energy gain is achieved solely through heat transfer to the working fluid by way of a heat exchanger.

Examples of IC engines

Internal-combustion engines are the most broadly applied and widely used power-generating devices currently in existence. Examples include gasoline (or spark-ignition [SI]) engines, diesel engines (sometimes referred to as compression-ignition [CI] engines), gas-turbine engines, and rocket propulsion systems.

The most common internal-combustion engine is the four-stroke gasoline-powered, homogeneous-charge, spark-ignition engine. This is because of its outstanding performance as a prime mover in the ground-transportation industry. Spark-ignition engines also are used in the aeronautics industry; however, aircraft gas turbines have become the prime movers in this sector due to the emphasis of the aeronautics industry on range, speed, and passenger comfort. The domain of internal-combustion engines also includes such exotic devices as supersonic combustion ramjet engines (scramjets), as typified by the space plane, and sophisticated rocket engines and motors, as those used on the U.S. Space Shuttle and other space vehicles.

It is the versatility and cost—both capital and operational—of conventional internal-combustion engines that have led to their widespread use in contemporary energy production. (C.L.P.II)

GASOLINE ENGINES

General characteristics. The gasoline engine is an intermittent-combustion engine. It is powered by the combustion of a premixed charge of air and gasoline, which is ignited electrically by a spark.

Most gasoline engines are of the so-called reciprocating piston type, but recent developments suggest that superior performance in some respects may be obtained from either rotary piston or turbine types (see below). Several terms are unique to the reciprocating piston engine. The piston–cylinder arrangement defines all terms relative to the size, location, and position of the piston within the cylinder (see Figure 18). Bore is the inner diameter of the cylinder. The volume at bottom dead centre (VBDC) is defined as the volume occupied between the cylinder head and the piston face when the piston is farthest from the cylinder head. The volume at top dead centre (VTDC) is that volume occupied when the piston is closest to the cylinder head; the distance between the piston face and cylinder head at VTDC is called the clearance. The distance traveled by the piston between its VTDC and VBDC locations is the stroke. The compression ratio of a reciprocating engine is

Basic terms

From R.A. Meyers (ed.), *Encyclopedia of Physical Science and Technology,* vol. 6 (1987), Academic Press, Inc.

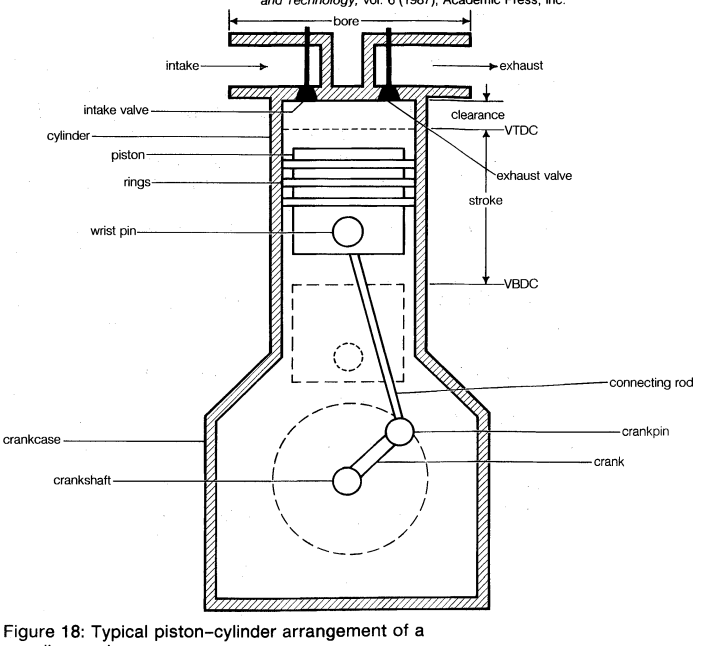


Figure 18: Typical piston–cylinder arrangement of a gasoline engine.

the ratio of VTDC to VBDC normalized to the VTDC value—*i.e.*, (VBDC/VTDC):1.

In all internal-combustion engines the products of combustion act directly on piston or rotor surfaces, whereas the external-combustion engine employs a secondary working fluid that is interposed between the combustion chamber and the power-producing elements. Fundamentally, the steam engine operates with a high-pressure working medium produced by utilizing the expansion accompanying the vaporization of a liquid (see above); by contrast, the internal-combustion engine utilizes the large volume of high-temperature combustion products that, when confined, become a high-pressure gaseous medium.

Combustion processes

Classification. The many types of internal-combustion engines can be grouped in a number of different ways on the basis of similarities among them. Important methods of classification include application, type of fuel and method of injection, ignition, reciprocating piston or rotary, cylinder arrangement, strokes per cycle, cooling system, and valve type and location. These various classifications will be discussed further as the various engine types are described.

Valve type and arrangement. Valves for controlling intake and exhaust may be located overhead, on one side, side and overhead, or on opposite sides of the cylinder. These are all the so-called poppet or mushroom valves consisting of a stem with one end enlarged to form a head that permits flow through a passage surrounding the stem when raised from its seat and prevents flow when the head is moved down to contact the valve seat formed in the cylinder block.

Another group of engines uses sliding valves that are usually of the sleeve type surrounding the cylinder bore.

Pressure application. Some power plants use the same combustion principle but apply the pressure resulting from combustion to different mechanical elements. There are, for example, gas-turbine engines in which the products of combustion are directed through nozzles against the blades of a turbine rotor to cause it to rotate. In the jet engine the products of combustion simply flow through a nozzle, and the reaction force tends to move the nozzle in the opposite direction.

Rotary engines

The Wankel and Tri-Dyne engines (see below) burn the fuel within the engine; they are rotary and do not have conventional cylinders fitted with reciprocating pistons. Instead, the gas pressure acts on surfaces formed by the configuration of a rotor. Both gas-turbine and jet engines have combustion furnaces separate from the power-producing units. The power is produced by the action of the products of combustion on the blades of the turbine or the interior wall of the jet nozzle (see *Gas-turbine engines* and *Jet engines* below).

Comparison with other engines. When the gasoline engine is compared with other types, certain similarities and differences as well as some advantages and disadvantages become apparent. The diesel engine and the gas engine (an engine utilizing a gas such as propane as the fuel) have a good deal in common with the gasoline engine, since they are all cylinder-and-piston engines that burn air–fuel mixtures in contact with moving components. The important difference that distinguishes the diesel engine is that it has no spark-ignition system. The diesel is heavier and more expensive per horsepower of output, but it has a longer life and operates at less cost per horsepower-hour because it burns less fuel. (For more specific details, see *Diesel engines* below.)

The gas engine

The gas engine has much in common with the gasoline engine; in fact, in some instances their differences are very slight at best. Structurally, the difference lies primarily in the substitution of a gas-mixing valve for a carburetor. The cylinder and piston configurations are the same. In general, gases have better antiknock qualities than gasoline (see below), permitting slightly higher compression ratios without knock or other combustion difficulties.

From the standpoint of application, the gas engine burning natural gas, manufactured gas, or industrial by-product gas is limited primarily to stationary power plant use because it must remain connected to the gas pipeline. If, however, the fuel is liquefied petroleum gas, sometimes called bottled gas, the containers of gas can be carried in a vehicle, leading to much flexibility in applications. The present obstacle is that facilities are not readily available for replenishing the gas supply. Dual carburetors have been produced experimentally that make it possible to operate an engine on either liquefied petroleum gas or gasoline; thus dual gas–gasoline engines are a distinct possibility.

Engine types. Of the different techniques for recovering the power from the combustion process the most important so far has been the four-stroke cycle, a conception now more than 100 years old.

Four-stroke cycle. The four-stroke cycle is illustrated in Figure 19. With the inlet valve open, the piston first descends on the intake stroke. An explosive mixture of gasoline vapour and air is drawn into the cylinder by the partial vacuum thus created. The mixture is compressed as the piston ascends on the compression stroke with both valves closed. As the end of the stroke is approached, the charge is ignited by an electric spark. The power stroke follows, with both valves still closed and the gas pressure, due to the expansion of the burned gas, pressing on the piston crown. During the exhaust stroke, the ascending piston forces the spent products of combustion through the open exhaust valve. The cycle then repeats itself. Each cycle thus requires four strokes of the piston—intake, compression, power, and exhaust—and two revolutions of the crankshaft.

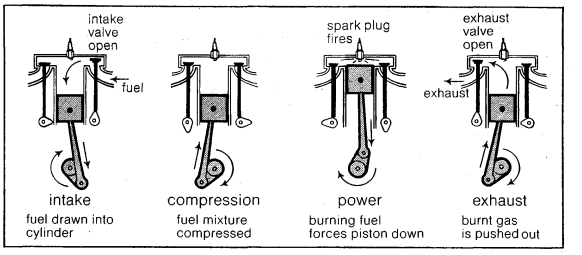


Figure 19: Strokes of the four-stroke cycle.

A disadvantage of the four-stroke cycle is that only half as many power strokes are completed as in the two-stroke cycle (see below) and only half as much power can be expected from an engine of a given size at a given operating speed. The four-stroke cycle, however, provides more positive clearing out of exhaust gases (scavenging) and reloading of the cylinders, reducing the amount of loss of fresh charge to the exhaust.

Two-stroke cycle. In the original two-stroke cycle (as developed in 1878), the compression and power stroke of the four-stroke cycle are carried out without the inlet and exhaust strokes, thus requiring only one revolution of the crankshaft to complete the cycle. Figure 20 illustrates the two-stroke-cycle engine of a so-called uniflow type in which the fresh fuel mixture is forced into the cylinder through circumferential ports by a rotary blower. The exhaust gases pass through poppet valves in the cylinder head that are opened and closed by a cam-follower mechanism. The valves are timed to begin opening toward the end of the power stroke after the cylinder pressure has dropped appreciably. The inlet ports in the cylinder wall start to uncover after the exhaust opening has decreased the cylinder pressure to the inlet pressure produced by the blower. The exhaust valves are allowed to remain open for a few

Uniflow two-stroke-cycle engine

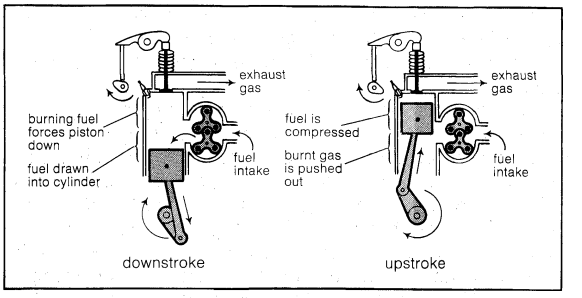


Figure 20: Blower-scavenged, two-stroke-cycle engine with uniflow scavenging.

degrees of crank rotation after the inlet ports have been covered by the rising piston on the compression stroke, thus allowing the persistency of flow more thoroughly to scavenge the cylinder. The compression and power strokes are similar to those of the four-stroke engine.

Crankcase compression

A simplified version of the two-stroke-cycle engine was developed some years later (introduced in 1891) by using crankcase compression to pump the fresh charge into the cylinder. Instead of intake ports extending entirely around the lower cylinder wall, this engine has intake ports only halfway around; a second set of ports starts a little higher in the cylinder wall in the other half of the cylinder bore. These larger ports lead to the exhaust system. The inlet ports connect to a transfer passage leading to the fully enclosed crankcase. A spring-loaded inlet valve admits air into the crankcase on the upward, or compression, stroke of the piston. Air trapped in the crankcase is compressed by the descent of the piston on its power stroke. The piston thus uncovers the exhaust ports near the end of the power stroke and slightly later it uncovers the inlet or transfer port on the opposite side of the cylinder to admit the compressed fresh mixture from the crankcase. The top face of the piston is designed to provide a deflector or baffle that directs the fresh load upward on the inlet side of the cylinder and then downward on the exhaust side, thus pushing the spent gases of the previous cycle out through the exhaust port on that side. This outflow continues after the inlet ports are covered by the rising piston on the compression stroke until the exhaust ports are covered and compression of the fresh load begins. This loading process, called loop scavenging, is the simplest known method of replacing the exhaust products with a fresh mixture and completing the cycle with only compression and power strokes.

Loop scavenging

Such a system is used in many small gasoline engines (*e.g.,* small outboard motors) and for gasoline-powered appliances. A disadvantage is that the return flow of the gases causes a slight loss of fresh charge through the exhaust ports. Because of this loss, carburetor engines operating on the two-stroke cycle lack the fuel economy of four-stroke engines. The loss can be avoided by equipping them with fuel-injection systems (see below) instead of carburetors and injecting the fuel directly into the cylinders after scavenging. Such an arrangement is attractive as a means of attaining high power output from a relatively small engine, and development of the turbocharger (see below *Supercharger*) for this application holds promise of further improvement.

Opposed-piston engine. The opposed-piston engine also provides uniflow scavenging. This engine (Figure 21A) has two pistons moving in opposite directions in the same cylinder. Two sets of ports extending entirely around the cylinder bore are so located that one set is covered and uncovered by one piston and the other set is controlled by the second piston. A second crankshaft, to which the upper pistons are attached, is located at the top of the engine and the two shafts are connected by gears.

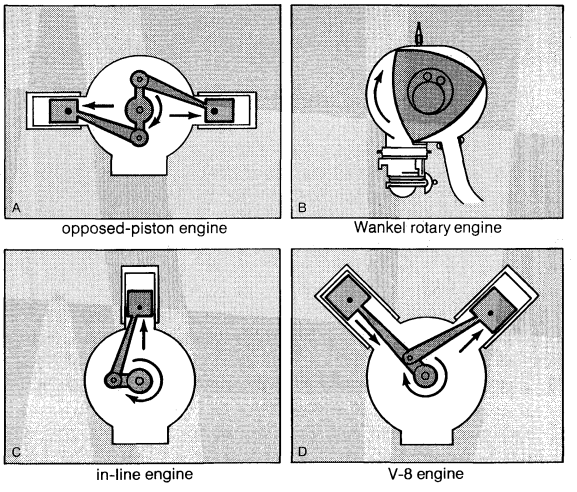


Figure 21: Certain types of gasoline engines.

The opposed-piston design has two major advantages: reciprocating masses move in opposite directions, providing excellent balance; and the poppet valves necessary in other uniflow-scavenged two-stroke-cycle engines are eliminated.

Wankel rotary engine. A rotary-piston internal-combustion engine developed in Germany is radically different in structure from conventional reciprocating piston engines. The engine was conceived by Felix Wankel, a specialist in the design of sealing devices, and experimental units were built and tested by a German firm beginning in 1956. Instead of pistons that move up and down in cylinders, the Wankel engine has an equilateral triangular orbiting rotor (see Figure 21B). The rotor turns in a closed chamber and the three apexes of the rotor maintain a continuous sliding contact with the curved inner surface of the casing. The curve-sided rotor forms three crescent-shaped chambers between its sides and the curved wall of the casing. The volumes of the chambers vary with the rotor motion. Maximum volume is attained in each chamber when the side of the rotor forming it is parallel with the minor diameter of the casing, and the volume is reduced to a minimum when the rotor side is parallel with the major diameter. Shallow pockets recessed in the flank of the rotor control the shape of the combustion chambers and establish the compression ratio of the engine.

Action of the rotor

In turning about its central axis the rotor must follow a circular orbit about the geometric centre of the casing. The necessary orbiting rotation is attained by means of a central bore in the rotor in which an internal gear is fitted to mesh with a stationary pinion fixed immovably to the centre of the casing. The rotor is guided by fitting its central bore to an eccentric formed on the output shaft that passes through the centre of the stationary pinion. This eccentric also harnesses the rotor to the shaft so that torque is applied when gas pressure is exerted against the rotor flanks as the fuel and air charges burn. A 3-to-1 gear ratio causes the output shaft to turn three times as fast as the rotor turns about the eccentric. Each quarter turn of the rotor completes an expansion or a compression, permitting intake, compression, expansion, and exhaust to be accomplished during one turn of the rotor. The only moving parts are the rotor and the output shaft.

The fuel mixture is supplied by a carburetor and enters the combustion chambers through an intake port in one of the end plates of the casing. An exhaust port is formed in one of the flattened sides of the casing wall and a spark plug is located in a pocket communicating with the chambers through a small throat in the opposite side of the casing wall.

The rotor and its gears and bearings are lubricated and cooled by oil circulating through the hollow rotor. The apex vanes are lubricated by a small amount of oil added to the fuel in proportions as low as 1 to 200. Water is circulated through cooling jackets in the casing, the entrance to which is located adjacent to the spark plug where the temperature tends to be highest.

Maintaining pressure-tight joints by suitable seals at the apexes and on the end faces of the rotor is a major design problem. Radial sliding vanes are fitted in slots at the three apex edges and kept in contact with the casing by expander springs. The end faces of the rotor are sealed by arc-shaped segmental rings fitted in grooves close to the curved edges of the rotor and pressed against the casing by flat springs.

Advantages of the Wankel engine

The major advantages of the Wankel engine are its small space requirements and low weight per horsepower, smooth and vibrationless operation, quiet operation, and low manufacturing costs resulting from mechanical simplicity. The absence of inertial forces from reciprocating parts and the elimination of spring-closed poppet valves permit operation at much higher speed than is practical for reciprocating piston engines, an advantage because shaft speed must be high for optimum performance. The induction of fresh fuel mixture and exhaust are more effective because the ports are opened and closed more rapidly than with poppet valves, and gas flow through them is almost continuous. Heat transfer and the resulting cooling requirement are low because the jacketed surface is small. Fuel economy is at least as good as that of conventional

gasoline engines, providing knock-free combustion with a wider range of permissible fuels. Lower weight and a lower centre of gravity make it much safer in an automobile in the event of a collision. There are approximately one third as many parts in a Wankel engine as in a typical six-cylinder automobile engine.

Tri-Dyne rotary engine. The Tri-Dyne engine, a British design, consists of three rotors (Figure 22). The large, triangular central rotor is called the power rotor. The other two are a combustion rotor and a barrier valve. The power rotor turns in the opposite direction from the combustion rotor and barrier valve. It has three curved lobes that fit into three semicircular cavities in the periphery of each of the two smaller rotors. The three are geared together by spur-shaped gears on the end of each rotor; all of them turn at the same speed. The motion is entirely rotary with no eccentricity. The three cavities in the combustion rotor form the combustion chambers and the profiles of all three rotors are such that, while not actually touching each other, they interact to connect these cavities alternately with the inlet and exhaust pipes and isolate them during the combustion process. It is not necessary that the cavities be positively sealed because of the high speed of operation. Clearance of 0.1 millimetre (0.004 inch) is provided between the interacting surfaces. Two spark plugs are installed in the casing at a point where they communicate with the combustion rotor cavities as they pass at the instant of firing. The advantage of the Tri-Dyne engine over the Wankel engine lies in the elimination of the seals that the latter requires at the apexes of its triangular rotor that limit the speed at which it can operate and that are difficult to lubricate.

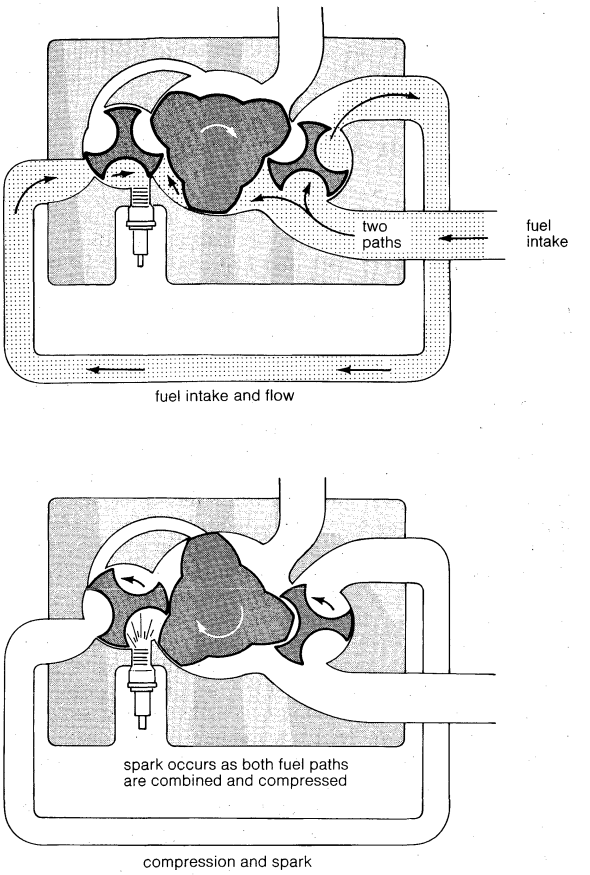


Figure 22: Simplified sketches showing operating principles of the Tri-Dyne engine.

Engine construction and operation. The overall structure of a gasoline engine depends almost entirely upon the intended application. Many components require only slight modification. Apart from the type of cycle (two- or four-stroke) the provision for mounting is the main structural difference among automotive, marine, stationary, and aviation engines. When a clutch and transmission are used, as in automobiles, the engine is commonly of the so-called unit-power-plant type with a bell-shaped housing surrounding the flywheel and attached to the rear flange of the cylinder block integral with, or attached to, the transmission gear case. The clutch is incorporated in the flywheel of the engine. Three-point suspension is used in such engines; that is to say, projections on each side of the bell housing fit into the vehicle side frame members and a central tubular extension at the centre of the front end of the cylinder block attaches to the front cross member of the frame. This construction permits some flexing of the vehicle frame without stressing the basic structure of the engine.

Unit-power-plant type

The following description of general engine construction indicates the essential components of an engine and introduces the nomenclature of the various parts. The four-stroke-cycle automobile engine is used as the basic type. Figure 23 shows a cross section of a typical automobile engine with the principal parts indicated.

Cylinder block. The main structural member of all automotive engines is a cylinder block that usually extends upward from the centre line of the main support for the crankshaft to the junction with the cylinder head. The block serves as the structural framework of the engine and carries the mounting pad by which the engine is supported in the chassis. Large, stationary power-plant engines and marine engines are built up from a foundation or bedplate and have upper and lower crankcases that are separate from the cylinder assemblies. The cylinder block of an automobile engine is a casting with appropriate machined surfaces and threaded holes for attaching the cylinder head, main bearings, oil pan, and other units.

Crankcase

The crankcase is formed by the portion of the cylinder block below the cylinder bores and the stamped metal oil pan that forms the lower enclosure of the engine and also serves as a lubricating oil reservoir or sump.

The cylinders are openings of circular cross section that extend through the upper portion of the block with interior walls bored and polished to form smooth, accurate bearing surfaces. The cylinders of heavy-duty engines are usually fitted with removable liners made of metal that is more wear-resistant than that used in the block casting.

There are two arrangements of cylinders in common automotive use—the vertical or in-line type (Figure 21C) and the V type (Figure 21D). The in-line engine has a single row of cylinders extending vertically upward from the crankcase and aligned with the crankshaft main bearings. The V type has two rows of cylinders, usually forming an angle of 60° or 90° between the two banks. V-8 engines (eight cylinders) are usually of the 90° type. Some small 6-cylinder aviation engines have horizontally opposed cylinders.

A passage bored lengthwise in the block houses the camshaft that operates the valves. A gear or chain compartment for the camshaft drive from the crankshaft is formed between the front or rear end of the block and a cover plate. The bell housing is formed at the rear of the cylinder block to enclose the flywheel and provide for

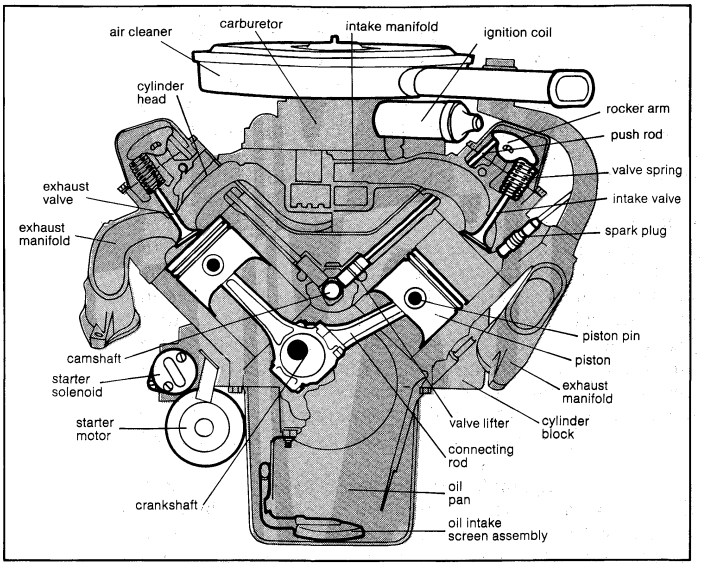


Figure 23: Cross section of a V-8 engine.
By courtesy of Cadillac Motor Car Division

attachment of a transmission housing. Water jackets are formed around the cylinders with suitable cored connecting passages for circulation of the coolant.

Block design

The design of the cylinder block is affected by the location of the valves of the four-stroke-cycle engine and by the provision of cylinder ports in the two-stroke type. An overhead-valve engine, which has largely replaced the L-head type, has its valves entirely in the cylinder head. The cylinder block of the L-head engine is extended to one side of the cylinder bores, with the valve seats and passages for inlet and exhaust, together with the valve guides, formed in this extension of the block. The cylinder head then becomes merely a water-jacketed cover, providing threaded locations for the spark plugs and with its underside so profiled that a combustion chamber of desired size and shape is formed above each cylinder bore. The shape of the space forming the combustion chamber when the piston is at its closest approach to the cylinder head and volume contained therein in relation to the piston displacement volume are extremely important in their effect on performance. The cylinder head of the valve-in-head engine is narrower and deeper and carries the valve seats, valve guides, and valve ports.

Combustion chamber. The size of the combustion chamber relative to the volume displaced by the piston establishes the compression ratio of the engine. The piston displacement is the volume swept by the piston during one stroke and is equal to the cross-sectional area of the cylinder multiplied by the length of the stroke. The larger volume above the piston at the lowest point in its stroke, divided by the combustion-chamber volume when the piston is at its highest point, is the compression ratio of the engine. The larger volume is the sum of the piston-displacement volume and the combustion-chamber volume. The compression ratio may thus be expressed as the ratio of the sum of the piston displacement volume and the combustion-chamber volume to the combustion-chamber volume. Compression ratio is the most important factor affecting the theoretical efficiency of the engine cycle. Because increasing the compression ratio is the best way to improve efficiency, compression ratios on automobile engines have tended to increase. This requires stronger, more durable materials.

Importance of compression ratio

Pistons. The pistons are cup-shaped cylindrical castings of steel or aluminum alloy. The upper, closed end, called the crown, forms the lower surface of the combustion chamber and receives the force applied by the combustion gases. The outer surface is machined to fit the cylinder bore closely and is grooved to receive piston rings that seal the gap between the piston and the cylinder wall. In the upper piston grooves there are plain compression rings that prevent the combustion gases from blowing past the piston. The lower rings are vented to distribute and limit the amount of lubricant on the cylinder wall. Piston pin supports (bosses) are cast in opposite sides of the piston and hardened steel pins fitted into these bosses pass through the upper end of the connecting rod.

Connecting rod and crankshaft. A forged steel connecting rod connects the piston to a throw (offset portion) of the crankshaft and converts the reciprocating motion of the piston to the rotating motion of the crank. The lower, larger end of the rod is bored to take a precision bearing insert lined with Babbitt or other bearing metal and closely fitted to the crankpin. V-type engines usually have opposite cylinders staggered sufficiently to permit the two connecting rods that operate on each crank throw to be side by side. Some larger engines employ fork-and-blade rods with the rods in the same plane and cylinders exactly opposite each other.

V-type engines

Each connecting rod in an in-line engine or each pair of rods in a V-type engine is attached to a throw of the crankshaft. Each throw consists of a crankpin with a bearing surface, on which the connecting rod bearing insert is fitted, and two radial cheeks that connect it to the portions of the crankshaft that turn in the main bearings, supported by the cylinder block. Sufficient throws are provided to serve all the cylinders, and the angles between them equal the angular firing intervals between the cylinders. The throws of a six-cylinder, four-stroke-cycle crankshaft are spaced 120° apart so that the six cylinders fire at equal intervals in two full rotations of the shaft. Those of an eight-cylinder engine are 90° apart. The position of each throw along the shaft depends upon the firing order of the cylinders. Firing sequence is chosen to distribute the power impulses along the length of the engine to minimize vibration. Consideration is also given to the fluid flow pattern in the intake and exhaust manifolds. The standard firing order for a six-cylinder engine is 1-5-3-6-2-4, which illustrates the practice of alternating successive impulses between the front and rear valves of the engine whenever possible. Balance is further improved by adding counterweights to the crankshaft to offset the eccentric masses of metal in the crank throws.

The crankshaft design also establishes the length of the piston stroke because the radial offset of each throw is equal to half the stroke imparted to the piston. The ratio of the piston stroke to the cylinder bore diameter is an important design consideration. In the early years of engine development, no logical basis for the establishment of this ratio existed, and a range from unity to 1½ was used by different manufacturers. As engine speeds increased, however, and it became apparent that friction horsepower increased with piston speed rather than with crankshaft rotating speed, there began a trend toward short-stroke engines. Strokes were shortened to as much as 20 percent less than the bores.

Crank-throw layout

From the requirement for the two-cylinder engine a general rule for the layout of the throws of four-stroke-cycle multicylinder crankshafts can be expressed. Regardless of the number of cylinders, two pistons must arrive at top dead centre (see above) in unison so that a second cylinder is ready to fire exactly 360° after each cylinder fires. Half of the cylinders then will fire during each turn of the crankshaft. To follow this rule, there must be an even number of cylinders in order that there may be pairs of cylinders whose pistons move in unison.

An eight-cylinder engine fires each time its crankshaft makes a quarter turn if the intervals between impulses are equal. The crankshaft for an eight-cylinder, in-line engine is designed with each of its eight throws a quarter turn away from another throw.

For best lengthwise balance, the cylinders whose pistons are in phase are the first and last cylinders of an in-line engine, the second and next to the last, continuing in that order with crank throws that are in alignment equidistant from the centre of the engine.

Valves, pushrods, and rocker arms. The valve-in-head engine has pushrods that extend upward from the cam followers to rocker arms mounted on the cylinder head that contact the valve stems and transmit the motion produced by the cam profile to the valves. Clearance (usually

termed tappet clearance) must be maintained between the ends of the valve stems and the lifter mechanism to assure proper closing of the valves when the engine temperature changes. This is done by providing pushrod length adjustment or by the use of hydraulic lifters.

Noisy and erratic valve operation can be eliminated with entirely mechanical valve lifter linkage only if the tappet clearance between the rocker arms and the valve stems is closely maintained at the specified value for the engine as measured with a thickness gauge. Hydraulic valve lifters, now commonly used on automobile engines, eliminate the need for periodic adjustment of clearance.

Hydraulic valve lifters

The hydraulic lifter comprises a cam follower that is moved up and down by contact with the cam profile, and an inner bore into which the valve lifter is closely fitted and retained by a spring clip. The valve lifter, in turn, is a cup closed at the top by a freely moving cylindrical plug that has a socket at the top to fit the lower end of the pushrod. This plug is pushed upward by a light spring that is merely capable of taking up the clearance between the valve stem and the rocker arm. A small hole is drilled in the bottom of the valve-lifter cup to admit lubricating oil that enters the cam follower from the engine lubricating system through a passage in the cylinder block. A small steel ball serves as a check valve to admit the oil into the valve-fitter cup but prevent its escape. When the clearance in the entire linkage between the cam profile and the valve stem is being taken up by the spring in the valve lifter, oil flows into the lifter chamber past the ball check and is trapped there to maintain this no-clearance condition as the engine operates. Expansion or contraction of the valve linkage is compensated by oil seepage from the lifter to correct for expansion of parts and oil flow into the chamber if clearance tends to be produced between the pushrod and the lifter. Complete closure of the valve is then assured at all times without tappet noise.

The intake valve must be open while the piston is descending on the intake stroke of the piston, and the exhaust valve must be open while the piston is rising on the exhaust stroke. It would seem, therefore, that the opening and closing of the two valves would occur at the appropriate top and bottom dead-centre points of the crankshaft. The time required for the valves to open and close, however, and the effects of high speed on the starting and stopping of the flow of the gases requires that for optimum performance the opening events occur before the crankshaft dead-centre positions and that the closing events be delayed until after dead centre.

Valve timing

All four valve events, inlet opening, inlet closing, exhaust opening, and exhaust closing, are accordingly displaced appreciably from the top and bottom dead centres. Opening events are earlier and closing events are later to permit ramps to be incorporated in the cam profiles to allow gradual initial opening and final closing to avoid slamming of the valves. Ramps are provided to start the lift gradually and to slow the valve down before it contacts its seat. Early opening and late closure are also for the purpose of using the inertia or persistence of flow of the gases to assist in filling and emptying the cylinder.

Camshaft. The camshaft, which opens and closes the valves, is driven from the crankshaft by a chain drive or gears on the front end of the engine. Because one turn of the camshaft completes the valve operation for an entire cycle of the engine and the four-stroke-cycle engine makes two crankshaft revolutions to complete one cycle, the camshaft turns half as fast as the crankshaft. It is located above and to one side of the crankshaft, which places it directly under the valves of the L-head engine or the pushrods that extend down from the rocker arms of the valve-in-head engine. Because of the long pushrods and the rocker arms, the speed of the valve-in-head engine is limited to that at which the cam followers can remain in contact with the cams when the valves are closing. Above that limiting speed the valves are said to float and their motion tends to become erratic. For this reason, the overhead-camshaft engine is increasing in popularity. Located immediately above the valves, this type of camshaft is driven either by a vertical shaft and bevel gears or by a cog belt.

Overhead camshafts

Flywheel. The cycle of the internal-combustion engine is such that torque (turning force) is applied only intermittently as each cylinder fires. Between these power impulses the pistons rising on compression and the opposition to rotation caused by the load carried by the engine apply negative torque. The alternating acceleration caused by the power impulse and deceleration caused by compression results in nonuniform rotation. To counter this tendency to slow down and speed up is the function of the flywheel, attached to one end of the crankshaft. The flywheel consists of a heavy circular cast-iron disk with a hub for attachment to the engine. Its heavy rotating mass has sufficient momentum to oppose all changes in its rotational speed and to force the crankshaft to turn steadily at this speed. The engine thus runs smoothly with no evidence of rotational pulsations. The outer rim of the flywheel usually carries gear teeth so as to mesh with the starter motor. The driving component of a clutch or fluid coupling for the transmission may be incorporated in the flywheel.

Bearings. The crankshaft has bearing surfaces on each crank throw and three or more main bearings. These are heavily loaded because of the reciprocating forces at each cylinder applied to the crankshaft and the weight of the crankshaft and flywheel. All but the smallest engines use split shell bearings, usually made of bronze with Babbitt-metal linings. The surface material is sufficiently soft to minimize the possibility of scoring the crankshaft in the event of inadequate lubrication. The smallest engines usually have cast Babbitt bearings. A small amount of bearing clearance is necessary to permit an oil film to separate the surfaces.

Ignition. Electric ignition systems may be classified as magneto and battery-and-coil systems. Although these are similar in basic principle, the magneto is self-contained and requires only the spark plugs and connecting wires to complete the system, whereas the battery-and-coil system involves several separate components. The circuit consists of a battery, one terminal of which is grounded while the other leads through a switch to the primary winding of the coil, and then to a circuit breaker where it is again grounded. Rotation of the circuit-breaker cam opens and closes the primary circuit. The secondary circuit, consisting of several thousand turns of fine wire, leads to the rotor of the distributor, which acts as a rotary switch, selecting the spark plug to be placed in the circuit. Each plug is connected to one of the outer terminals of the distributor to receive an electrical impulse in proper sequence. When the primary circuit is broken, a high potential (up to 20,000 volts) is developed in the secondary winding and conducted to the appropriate spark plug.

Spark plug

The spark plug is an important component of the ignition system and is the one that must operate under the most severe conditions. Because it is exposed to combustion-chamber temperatures and pressures and contaminating products of combustion, it requires more service attention and is usually the shortest-lived component of the gasoline engine. It consists of a steel shell threaded to fit a standard 14-millimetre hole in the cylinder head. A copper gasket insures a gastight fit between cylinder head and plug. A fused ceramic insulating element is molded into the plug body and the steel centre electrode passes through the insulator up to the connector to which the high-voltage lead from the distributor is attached. The other electrode is welded to the metal body of the plug, which is grounded to the cylinder head.

It is essential that the spark gap be as specified for the particular engine. Gauges are available to aid in making this adjustment by bending the ground electrode as required. Manufacturers specify gaps ranging from 0.508 to 1.016 millimetres between the centre electrode and the ground electrode. If the plug gap is too large, the possibility of misfiring increases. If the gap is too small, the spark will not be sufficiently intense. Gap growth from erosion of the electrodes may be corrected. The high voltage for the spark plug may also be produced by a capacitor discharge ignition system. Such a system consists of a source of 250 to 300 volts direct-current power applied to a storage capacitor, a device for storing an electric charge.

Capacitor ignition system

A lead from the capacitor goes to one side of the spark coil primary through cam-actuated breaker points or an electronic switching device. At the instant this switching device establishes a contact, the capacitor discharges through the primary of the spark coil and an instantaneous high voltage is delivered to the distributor and thence to the spark plug.

The capacitor discharge system stem provides a more intense spark, thus improving starting a cold or flooded engine. It continues to fire the plugs when they are fouled by carbon or other deposits or when the spark gap has widened because of erosion of the points. Other notable advantages include increased spark plug life, improved firing over a wider speed range, and better moisture tolerance.

A magneto is a fixed-magnet, alternating-current generator designed to generate sufficient voltage to fire the spark plugs. A high-tension magneto is entirely self-contained and requires only spark plugs, wires, and switches to do what is required in meeting ignition requirements.

Carburetor. The gasoline carburetor is a device that introduces fuel into the air stream as it flows into the engine. A simple carburetor is shown diagrammatically in Figure 24. Gasoline is maintained in the float chamber by the float-actuated valve at a level slightly below the outlet of the jet. Air flows downward through the throat, past the throttle valve, and into the intake manifold. A throat is formed by the reduced diameter, and acceleration of the air through this smaller passage causes a decrease in pressure proportional to the amount of air flowing. This decrease in throat pressure results in fuel flow from the jet into the air stream. Any increase in air flow caused by change in engine speed or throttle position increases the pressure differential acting on the fuel and causes more fuel to flow.

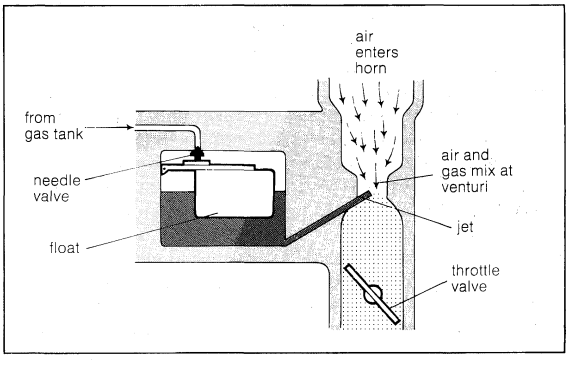


Figure 24: A simple carburetor.

The volume ratio of air to fuel established by the throat and fuel jet sizes will be maintained with increased flow, but the weight of fuel per kilogram (or pound) of air increases because the air expands to a lower density as the throat pressure decreases. This enriching tendency necessitates the inclusion of a compensating device in a practical carburetor. Carburetor design is further complicated by the need for an enriching device to provide a maximum-power ratio at full throttle, a choke to facilitate starting a cold engine, an idling system to provide the special needs of light-load operation, and an accelerating device to supply additional fuel while the throttle is being opened.

Fuel injection. Gasoline-injection systems in which the fuel is forcefully injected into the cylinder by a pump were available for airplane engines before World War II and were extensively used then in aircraft. The performance of engines with such equipment was excellent, but the much greater cost of fuel-injection systems compared with that of carburetors limited their application.

The above-mentioned form of fuel preparation is termed cylinder-head injection. Such a system, employed in stratified-charge engines, involves the injection of fuel directly into each cylinder under high pressure. It is well-suited for this type of engine, which is designed to permit operation under very fuel-lean conditions. Examples of stratified engines include the divided-chamber, axially stratified-charge, and direct-injection varieties (Figure 25).

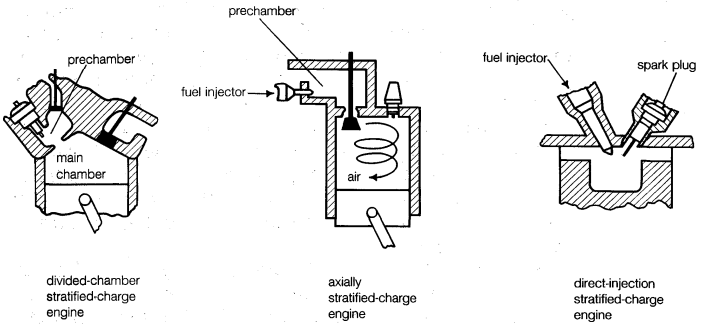


Figure 25: Stratified-charge engine design.

From R.A. Meyers (ed.), *Encyclopedia of Physical Science and Technology*, vol 6 (1987), Academic Press, Inc.

Efforts to simplify and lower the cost of fuel-injection equipment without impairing performance have yielded multicylinder pumps that can compete in cost with four-barrel carburetors. Gasoline-injection equipment may consist of distributor systems employing a single pump for all of the cylinders, or multipumps.

Advantages of fuel injection

The principal advantages of gasoline injection over carburetors are improved fuel economy because of more accurate fuel and air proportioning, greater power because of the elimination of fuel heating, elimination of inlet icing, and more uniform and direct delivery of fuel load to the cylinders.

Supercharger. The efficiency of the charging process in an automotive engine usually rises to a peak of slightly more than 80 percent at about half the rated speed of the engine and then decreases considerably at higher speed. This change in air charge per cycle with engine speed is reflected in proportionate changes in the torque, or turning effort, applied to the crankshaft and causes the power that the engine can deliver at full throttle to reach a maximum as engine speed increases. At speeds above this peaking speed, the air charge introduced per cycle falls off so rapidly that less power is developed than at lower speeds. The inability of the engine to draw in a full charge of fresh air at high speeds limits the power output of the engine.

Supercharging overcomes this disadvantage by the use of a pump or blower to raise the pressure of the air supplied to the cylinders, increasing the weight of charge. The loss in power suffered by unsupercharged engines at high altitudes can be largely restored; it is also possible to more than double the power of an engine by supercharging. Increased charge density and temperature, resulting from supercharging, increase the tendency for combustion knock or roughness in the spark-ignition engine and thus necessitates an undesirable decrease in compression ratio or the use of an antiknock fuel.

Turbocharger

The supercharging blower may be geared to the crankshaft, in which case the power consumed in driving it is added to the friction loss of the engine. A turbocharger employs a gas turbine operated by the exhaust gases to drive a centrifugal blower. The turbocharged engine not only gains increased power capacity but also operates at improved fuel economy. Airplane engines are usually supercharged both by geared blowers and by turbochargers to provide the large pumping capacity needed at high altitude.

Since compressing air prior to introducing it into the cylinder increases the charge-air temperature, the mass of air that can be introduced into the engine is less than that which would be possible if the compressed air were at ambient temperature. Consequently, engine charge-air coolers, commonly referred to as either intercoolers or aftercoolers, are used to reduce the temperature of the charge air. Both air-to-coolant and air-to-air type coolers are available.

Cooling system. The cylinders of internal-combustion engines require cooling because of the inability of the engine to convert all of the energy released by combustion into useful work. Liquid cooling is employed in most gasoline engines, whether the engines are for use in automobiles or elsewhere. The liquid is circulated around

the cylinders to pick up heat and then through a radiator to dissipate the heat. Usually a thermostat is located in the circulating system to maintain the design jacket temperature—71° to 82° C. The cooling system is usually pressurized to raise the boiling point of the coolant so that a higher outlet temperature can be maintained to improve thermal efficiency and increase the heat transfer capacity of the radiator. A pressure cap on the radiator maintains this pressure by valves that open outwardly at the design pressure and inwardly to prevent a vacuum as the system cools.

Some engines, particularly aviation engines and small units for mowers, chain saws, and other tools, are air cooled. Air cooling is accomplished by forming thin metal fins on the exterior surfaces of the cylinders to increase the rate of heat transfer by exposing more metal surface to the cooling air. Air is forced to flow rapidly through the spaces between the fins by ducting air toward the engine.

Air cooling

Lubrication system. Lubrication is employed to reduce friction by interposing a film between rubbing parts. The lubrication system must continuously replace the films.

The lubricants commonly employed are refined from crude oil after the fuels have been removed. Their viscosities must be appropriate for each engine and the oil must be suitable for the severity of the operating conditions. Oils are improved with additives that reduce oxidation, inhibit corrosion, and act as detergents to disperse deposit-forming gums and solid contaminants. Various systems of numbers are used to designate oil viscosity; the lower the number, the lighter the body of the oil. Certain oils contain additives that oppose their change in viscosity between the winter and summer.

Oil filters, if regularly serviced, can remove solid contaminants from crankcase oil, but chemical reactions may form liquids that are corrosive and damaging. Depletion of the additives also limits the useful life of lubricating oils.

The lubrication system is fed by the oil sump that forms the lower enclosure of the engine. Oil is taken from the sump by a pump, usually of the gear type, and delivered under pressure to a system of passages or channels drilled through the engine. In some instances a so-called full-flow filter runs the length of the engine between the pump and the main oil passage. In other engines bypass filters continuously bleed off a small quantity of oil and return the filtered oil to the sump.

Oil is supplied under pressure to crankshaft and camshaft main bearings. Adjacent crank throws are drilled to enable the oil to flow from the supply at the main bearings to the crankpins. Leaking oil from all of the crankshaft bearings is sprayed on the cylinder walls, cams, and up into the pistons to lubricate the piston pins. Additional passages intersect the cam-follower openings and supply oil to hydraulic valve lifters when used. A spring-loaded pressure-relief valve maintains the pressure at the proper level.

Exhaust system. Exhaust gases from an internal-combustion engine are passed through a muffler to suppress audible vibrations. When the exhaust valve opens, the pressure in the engine causes an initial gas outflow at explosive velocity. Successive discharges from the cylinders set up pressure pulsations that produce a sharp barking sound. The muffler damps out or absorbs these pulsations so that the gases leave the outlet as a relatively smooth, quiet stream.

Mufflers of early design contained sets of baffles that reversed the flow of the gases or otherwise caused them to follow devious paths so that interference between the pressure waves reduced the pulsations. The mufflers most commonly used in modern motor vehicles employ resonating chambers connected to the passages through which the gases flow. Gas vibrations are set up in each of these chambers at the fundamental frequency determined by its dimensions. These vibrations cancel or absorb those present in the exhaust stream of about the same frequency. Several such chambers, each tuned to one of the predominant frequencies present in the exhaust stream, effectively reduce noise.

Emission control devices for reducing air pollution are added to the exhaust system. Beds of a suitable catalyst (a material for promoting desirable reactions) are placed in a mufflerlike chamber to reduce unburned hydrocarbons, carbon monoxide, and, in some instances, nitrogen oxides in the exhaust output. This device, called the catalytic converter, is used in conjunction with various other kinds of emission control systems.

Emission control

The reactor system for controlling emission is composed of a belt-driven air compressor connected to small nozzles installed in the exhaust manifold facing the outlet from each exhaust valve. A small jet of air is thus directed toward the red-hot outflowing combustion products to provide oxygen to consume the hydrocarbons and carbon monoxide.

Fuels. Gasoline was originally considered dangerous and was discarded and destroyed at early refineries, which were manufacturing kerosene for lamps. As the gasoline engine developed, gasoline and the engine were harmonized to attain the best possible matching of characteristics. The most important properties of gasoline are its volatility and antiknock quality. Volatility is a measure of the ease of vaporization of gasoline in the carburetor.

Volatility and antiknock quality

To suit the needs of a modern engine a gasoline must have the volatility for which the fuel system of the engine was designed and antiknock quality sufficient to avoid knock under normal operation. Although other specifications must also be met, volatility and knock rating are the most important. The size and structural arrangement of the molecules principally determine the knocking tendency of a gasoline as well as its volatility.

Tetraethyl lead, added to gasolines for many years to improve antiknock fueling, has been found to contaminate the exhaust gases with poisonous lead oxides, and the practice is ending. Lower compression ratios and improved combustion-chamber designs are eliminating the need for extremely high antiknock gasolines.

Lubricating oil is added to gasoline used in crankcase-compression two-stroke cycle engines.

Performance. The performance of an engine is expressed in terms of power, speed, and fuel economy. The three quantities are evaluated with a dynamometer, a laboratory device that applies a controllable load in the form of resistance to the turning of the crankshaft and also measures the torque exerted at the shaft coupling. The resistance imposed by a dynamometer may be so adjusted that the desired engine speed is established at any throttle position. It is thus possible to run the engine at various speeds throughout its operating range, to maintain these operating conditions continuously, and to measure the precise load and speed at which each run is made. Additional test equipment permits measurement of the exact quantity of fuel consumed as well as the duration of the runs. From these data the power–speed–economy relationships can be calculated and performance plotted.

The power produced by an engine is, as explained earlier, expressed in horsepower. When the power developed is measured by means of a dynamometer or similar braking device, it is called brake horsepower. This is the power actually delivered by the engine and is therefore the capacity of the engine. The power developed in the combustion chambers of the engine is greater than the delivered power because of friction and other mechanical losses. This power loss, called the friction horsepower, can be evaluated by "motoring" the engine (driving it in a forward direction) with a suitable dynamometer when no fuel is being burned. The power developed in the cylinder can then be found by adding the friction horsepower to the brake horsepower. This quantity is the indicated horsepower of the engine, so called from an instrument known as the engine indicator, which is used to measure the pressure on the piston and thus calculate the power developed in the cylinder.

Power measurements

Mechanical efficiency is defined as brake horsepower in percent of indicated horsepower and is usually between 70 percent and 90 percent within the normal operating speed range.

A quantity called brake mean effective pressure is obtained by multiplying the mean effective pressure of an engine by its mechanical efficiency. This is a commonly used index expressing the ability of the engine, per unit of cylinder bore, to develop useful pressure in the cylinders

and delivery power. If the power delivered is increased by any change other than an increase in speed or cylinder dimensions, its brake mean effective pressure increases proportionately.

Applications. Gasoline engines can be built to meet the requirements of practically any conceivable power-plant application. In some instances, however, other kinds of engines or electric motors have certain advantages. The important applications for which the gasoline engine is most likely to be chosen in preference to other types are in the areas of passenger automobiles, small trucks and buses, aircraft, outboard and small inboard marine units, moderate-sized stationary pumping, lighting plant, machine tool and similar installations, and power tools.

Development of gasoline engines. While attempts to devise heat engines were made in ancient times, the steam engine of the 18th century was the first successful type. The internal-combustion engine, which followed in the 19th century as an improvement over the steam engine for many applications, cannot be attributed to any single inventor. The piston, thought to date as far back as 150 BC, was used by metalworkers in pumps for blowing air. The piston (and cylinder) was basic to the steam engine, which brought the component to a high state of efficiency. The steam engine, however, suffered from low thermal efficiency, great weight and bulk, and inconvenience of operation, all of which were primarily traceable to the necessity of burning the fuel in a furnace separate from the engine. It became evident that a self-contained power unit was desirable.

Early use of the piston

As early as the 17th century, several experimenters first tried to use hot gaseous products to operate pumps. By 1820 an engine was built in England in which hydrogen–air mixtures were exploded in a chamber. The chamber was then cooled to create a vacuum acting on a piston. The sale of such gas engines began in 1823. They were heavy and crude but contained many essential elements of later, more successful devices. In 1824 the French engineer Sadi Carnot published his now classic pamphlet "Reflections on the Motive Power of Heat," which outlined fundamental internal-combustion theory. Over the next several decades inventors and engineers built engines that used pressure produced by the combustion of fuels rather than a vacuum and engines in which the fuel was compressed before burning. None of them succeeded in developing an operational system, however. Finally, in 1860 Étienne Lenoir of France marketed an engine that operated on illuminating gas and provided reasonably satisfactory service. The Lenoir engine was essentially a converted double-acting steam engine with slide valves for admitting gas and air and for discharging exhaust products. Although the Lenoir engine developed little power and utilized only about 4 percent of the energy in the fuel, hundreds of these devices were in use in France and Britain within five years. They were used for powering water pumps and printing presses and for completing certain other tasks that required only limited power output.

The ideal operating cycle

A major theoretical advance occurred with the publication in 1862 of a description of the ideal operating cycle of an internal-combustion engine. The author, the French engineer Alphonse Beau de Rochas, laid down the following conditions as necessary for optimum efficiency: maximum cylinder volume with minimum cooling surface, maximum rapidity of expansion, maximum ratio of expansion, and maximum pressure of the ignited charge. He described the required sequence of operations as (1) suction during an entire outstroke of the piston, (2) compression during the following instroke, (3) ignition of the charge at dead centre and expansion during the next outstroke (the power stroke), and (4) expulsion of the burned gases during the next instroke. The engine Beau de Rochas described thus had a four-stroke cycle, in contrast to the two-stroke cycle (intake-ignition and power-exhaust) of the Lenoir engine. Beau de Rochas never built his engine, and no four-stroke engine appeared for more than a decade. Finally in 1876 the German engineer Nikolaus A. Otto built an internal-combustion unit based on Beau de Rochas's principle. (Otto's firm, Otto and Langen, had produced and marketed an improved two-stroke engine several years earlier.) The four-stroke Otto engine was an immediate success. In spite of its great weight and poor economy, nearly 50,000 engines with a combined capacity of about 200,000 horsepower were sold in 17 years, followed by the rapid development of a wide variety of engines of the same type. Manufacture of the Otto engine in the United States began in 1878, following the grant to Otto of a U.S. patent in 1877.

The Otto engine

Eight years later Gottlieb Daimler and Wilhelm Maybach, former associates of Otto, developed the first successful high-speed four-stroke engine and invented a carburetor that made it possible to use gasoline for fuel. They employed their engine to power a bicycle (perhaps the world's first motorcycle) and later a four-wheeled carriage. At about the same time, another German mechanical engineer, Carl Benz, built a one-cylinder gasoline engine to power what is often considered the first practical automobile. The engines built by Daimler and Benz were fundamentally the same as today's basic gasoline engine. For information about subsequent enhancements and advances, see TRANSPORTATION: *Modern automotive systems.* (O.C.C./C.L.P.II)

DIESEL ENGINES

General characteristics. The diesel engine is an intermittent-combustion piston–cylinder device. It operates as either a two-stroke or four-stroke cycle (see *Gasoline engines* above); however, unlike the spark-ignition engine, the diesel engine induces only air into the combustion chamber on its intake stroke. The air is heated as compression occurs during the compression stroke. Diesel engines are typically constructed with compression ratios in the range 14:1 to 22:1. Both two-stroke and four-stroke engine designs can be found among engines with bores (cylinder diameters) less than 600 millimetres. Engines with bores of greater than 600 millimetres are almost exclusively two-stroke cycle systems.

The diesel engine gains its energy by burning fuel injected or sprayed into the compressed, hot air charge within the cylinder. The air must be heated to a temperature greater than the temperature at which the injected fuel can ignite. Fuel sprayed into air that has a temperature higher than the "auto-ignition" temperature of the fuel spontaneously reacts with the oxygen in the air and burns. Air temperatures are typically in excess of 526° C; however, at engine start-up, supplemental heating of the cylinders is usually required, since the temperature of the air within the cylinders is determined by both the engine's compression ratio and its current operating temperature. Diesel engines are sometimes called compression-ignition engines because initiation of combustion relies on air heated by compression rather than on an electric spark.

Ignition by compression of air charge

In a diesel engine, fuel is introduced as the piston approaches the top dead centre of its stroke (see *Gasoline engines* above). The fuel is introduced under high pressure either into a precombustion chamber (Figure 26) or directly into the piston–cylinder combustion chamber. With the exception of small, high-speed systems, diesel engines use direct injection.

Direct injection

Diesel-engine fuel-injection systems are typically designed to provide injection pressures in the range of seven to 70 megapascals (1,000 to 10,000 pounds per square inch). There are, however, a few higher pressure systems.

Precise control of fuel injection is critical to the performance of a diesel engine. Since the entire combustion process is controlled by fuel injection, injection must begin at the correct piston position (*i.e.,* crank angle). At first, the fuel is burned in nearly a constant-volume process while the piston is near top dead centre. As the piston moves away from this position, fuel injection is continued, and the combustion process then appears as a nearly constant-pressure process.

The combustion process in a diesel engine is heterogeneous—that is to say, the fuel and air are not premixed prior to initiation of combustion. Consequently, rapid vaporization and mixing of fuel in air is very important to thorough burning of the injected fuel. This places much emphasis on injector nozzle design, especially in direct-injection engines.

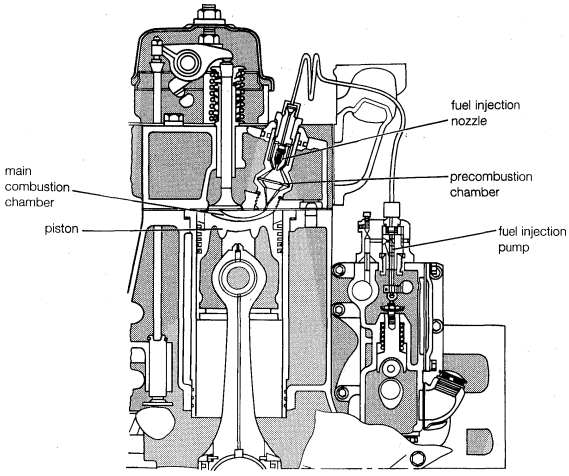


Figure 26: Diesel engine equipped with a precombustion chamber.

From H.A. Sorensen, *Energy Conversion Systems*, copyright © 1983 by John Wiley & Sons, Inc.; reprinted by permission of John Wiley & Sons, Inc.

Engine work is obtained during the power stroke. The power stroke includes both the constant-pressure process during combustion and the expansion of the hot products of combustion after fuel injection ceases.

Diesel engines are often turbocharged and aftercooled (see *Supercharger* above). Addition of a turbocharger and aftercooler can enhance the performance of a diesel engine both in terms of power and efficiency.

High efficiency of diesel engines

The most outstanding feature of the diesel engine is its efficiency. By compressing air, rather than using an air–fuel mixture, the diesel engine is not limited by the preignition problems that plague high-compression spark-ignition engines. Thus higher compression ratios can be achieved with diesel engines than with the spark-ignition variety; commensurately, higher theoretical cycle efficiencies, when compared to the latter, can often be realized. It should be noted that for a given compression ratio, the theoretical efficiency of the spark-ignition engine is greater than that of the compression-ignition engine; however, in practice, it is possible to operate compression-ignition engines at compression ratios high enough to produce efficiencies greater than those attainable with spark-ignition systems. Furthermore, diesel engines do not rely on throttling the intake mixture to control power. As such, the idling and reduced power efficiency of the diesel is far superior to that of the spark-ignition engine.

High pollutant emissions

The principal drawback of diesel engines is their emission of air pollutants. These engines typically discharge high levels of particulate matter (soot), reactive nitrogen compounds (commonly designated NO_x), and odour compared to spark-ignition engines. Consequently, in the small engine category (see below), consumer acceptance is low.

Major types of diesel engines. There are three basic size groups of diesel engines based on power—namely, small, medium, and large. The small engines have power output values of less than 188 kilowatts, or 252 horsepower. This is the most commonly produced diesel-engine type. These engines are used in automobiles, light trucks, and some agricultural and construction applications, and as small stationary electrical power generators (such as those on pleasure craft) and as mechanical drives. They are typically direct-injection, in-line, four- or six-cylinder engines. Many are turbocharged with aftercoolers.

Medium engines have power capacities ranging from 188 to 750 kilowatts, or 252 to 1,006 horsepower. The majority of these engines are used in heavy-duty trucks (those of the class 6, 7, and 8 variety). They are usually direct-injection, in-line, six-cylinder turbocharged/aftercooled engines. Some V-8 and V-12 engines also belong to this size group.

Large diesel engines have power ratings in excess of 750 kilowatts. These unique engines are used for marine, locomotive, and mechanical drive applications and for electrical power generation. In most cases, they are direct-injection, turbocharged/aftercooled systems. They may operate at as low as 500 revolutions per minute when good reliability and durability are critical.

Engine structure and components. As noted earlier, diesel engines are designed to operate on either the two- or four-stroke cycle. In the typical four-stroke-cycle engine (Figure 27), the intake and exhaust valves and the fuel-injection nozzle are located in the cylinder head. Often dual valve arrangements, two intake and two exhaust valves, are employed.

Dual valve arrangement

Use of the two-stroke cycle can eliminate the need for one or both valves in the engine design. Scavenging and intake air is usually provided through ports in the cylinder liner. Exhaust can be either through valves located in the cylinder head or through ports in the cylinder liner. Engine construction is simplified when using a port design instead of one requiring exhaust valves.

Diesel engine starting. A diesel engine is started by driving it from some external power source until conditions have been established under which the engine can run by its own power. The most positive starting method is by admitting air at about 1.7 to nearly 2.4 megapascals to each of the cylinders in turn on their normal firing stroke. The compressed air becomes heated sufficiently to ignite the fuel. Other starting methods involve auxiliary equipment and include admitting blasts of compressed air to an air-activated motor geared to rotate a large engine's flywheel; supplying electric current to an electric starting motor, similarly geared to the engine flywheel; or by means of a small gasoline engine geared to the engine flywheel. The selection of the most suitable starting method depends on the physical size of the engine to be started, the nature of the connected load, and whether or not the load can be disconnected during starting.

Fuel for diesels. Petroleum products normally used as fuel for diesel engines are distillates composed of heavy hydrocarbons, with at least 12 to 16 carbon atoms per molecule. These heavier distillates are taken from crude oil after the more volatile portions used in gasoline are removed. The boiling points of these heavier distillates range from 177° to 343° C. Thus, their evaporation temperature is much higher than that of gasoline, which has fewer carbon atoms per molecule. Specifications for diesel fuels published in 1970 listed three grades: the first was a volatile distillate recommended for high-speed engines with frequent and wide variations in load and speed; the second, a distillate for high-speed engines in services with high loads and uniform speeds; and the third, a fuel for low- and medium-speed engines in service with sustained loads.

Distillates composed of heavy hydrocarbons

Water and sediment in fuels can be harmful to engine operation; clean fuel is essential to efficient injection systems. Fuels with a high carbon residue can be handled best by engines of low-speed rotation. The same applies to those with high ash and sulfur content. The cetane number, which defines the ignition quality of a fuel, is ascertained by adjusting a mixture of cetane and alpha-methyl-naphthalene until it has the same ignition quality as the fuel being tested. The percentage of cetane in this mixture is then the cetane number of the fuel under test. For the first two grades of diesel fuel described above, the minimum cetane number is 40; for the third grade, the minimum is 30, representing 30 percent cetane in the fuel.

Development of diesel engines. *Early work.* Rudolf Diesel, a German engineer, conceived the idea for the

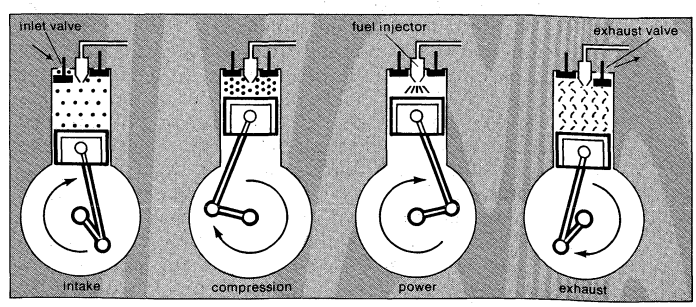


Figure 27: *Four-stroke diesel engine.*
The sequence of cycle events is shown here.

engine that now bears his name after seeking a device to increase the efficiency of the Otto engine (see *Development of gasoline engines* above). Diesel realized that the electric ignition process of the gasoline engine could be eliminated if, during the compression stroke of a piston–cylinder device, compression could heat air to a temperature higher than the auto-ignition temperature of a given fuel. Diesel proposed such a cycle in his patents of 1892 and 1893.

Originally, either powdered coal or liquid petroleum was proposed as fuel. Diesel saw powdered coal, a by-product of the Saar coal mines, as a readily available fuel. Compressed air was to be used to introduce coal dust into the engine cylinder; however, controlling the rate of coal injection was difficult, and after the experimental engine was destroyed by an explosion, Diesel turned to liquid petroleum. He continued to introduce the fuel into the engine with compressed air.

First commercial diesel engine

The first commercial engine built on Diesel's patents was installed in St. Louis, Mo., by Adolphus Busch, a brewer who had seen one on display at an exposition in Munich and had purchased a license from Diesel for the manufacture and sale of the engine in the United States and Canada. The engine operated successfully for years and was the forerunner of the Busch–Sulzer engine that powered many submarines of the U.S. Navy in World War I. Another diesel engine used for the same purpose was the Nelseco, built by the New London Ship and Engine Company in Groton, Conn.

The diesel engine became the primary power plant for submarines during World War I. It was not only economical in the use of fuel but also proved reliable under wartime conditions. Diesel fuel, less volatile than gasoline, was more easily stored and handled.

At the end of the war many men who had operated diesels were looking for peacetime jobs. Manufacturers began to adapt diesels for the peacetime economy. One modification was the development of the so-called semidiesel that operated on a two-stroke cycle at a lower compression pressure and made use of a hot bulb or tube to ignite the fuel charge. These changes resulted in an engine less expensive to build and maintain.

Fuel-injection technology. One objectionable feature of the full diesel was the necessity of a high-pressure, injection air compressor. Not only was energy required to drive the air compressor, but the sudden expansion of the air compressed to 6.9 megapascals when it entered the cylinder in which the pressure was only about 3.4 to 4.1 megapascals resulted in a refrigerating effect that delayed ignition. Diesel had needed high-pressure air with which to introduce powdered coal into the cylinder; when liquid petroleum replaced powdered coal as fuel, a pump could be made to take the place of the high-pressure air compressor.

Substitution of pumps for air compressors

There were a number of ways in which a pump could be used. In England the Vickers Company used what was called the common-rail method, in which a battery of pumps maintained the fuel under pressure in a pipe running the length of the engine with leads to each cylinder. From this rail (or pipe) fuel-supply line, a series of injection valves admitted the fuel charge to each cylinder at the right point in its cycle. Another method employed cam-operated jerk, or plunger-type, pumps, to deliver fuel under momentarily high pressure to the injection valve of each cylinder at the right time.

The elimination of the injection air compressor was a step in the right direction, but there was yet another problem to be solved: the engine exhaust contained an excessive amount of smoke, even at outputs well within the horsepower rating of the engine and even though there was enough air in the cylinder to burn the fuel charge without leaving a discoloured exhaust that normally indicated overload. Engineers finally realized that the problem was that the momentarily high-pressure injection air exploding into the engine cylinder had diffused the fuel charge more efficiently than the substitute mechanical fuel nozzles were able to do, with the result that without the air compressor, the fuel had to search out the oxygen atoms to complete the combustion process, and since oxygen makes up only 20 percent of the air, each atom of fuel had only one chance in five of encountering an atom of oxygen. The result was improper burning of the fuel.

The usual design of a fuel-injection nozzle introduced the fuel into the cylinder in the form of a cone spray, with the vapour radiating from the nozzle, rather than in a stream or jet. Very little could be done to diffuse the fuel more thoroughly. Improved mixing had to be accomplished by imparting additional motion to the air, most commonly by induction-produced air swirls or a radial movement of the air, called squish, or both, from the outer edge of the piston toward the centre. Various methods have been employed to create this swirl and squish. Best results are apparently obtained when the air swirl bears a definite relation to the fuel-injection rate. Efficient utilization of the air within the cylinder demands a rotational velocity that causes the entrapped air to move continuously from one spray to the next during the injection period, without extreme subsidence between cycles.

Swirl or squish

Price's engine. In 1914 a young American engineer, William T. Price, began to experiment with an engine that would operate with a lower compression ratio than that of the diesel and at the same time would not require either hot bulbs or tubes. As soon as his experiments began to show promise, he applied for patents.

In Price's engine the selected compression pressure of nearly 1.4 megapascals did not provide a high enough temperature to ignite the fuel charge when starting. Ignition was accomplished by a fine wire coil in the combustion chamber. Nichrome wire was used for this because it could easily be heated to incandescence when an electric current was passed through it. The experimental engine had a single horizontal cylinder with a bore of 43 centimetres and a stroke (maximum piston movement) of 48 centimetres and operated at 257 revolutions per minute. Because the nichrome wire required frequent replacement, the compression pressure was raised to 2.4 megapascals, which did provide a temperature high enough for ignition when starting. Some of the fuel charge was injected before the end of the compression stroke in an effort to increase the cycle timing and to keep the nichrome wire glowing hot.

In the meantime many engines of the two-stroke cycle, semidiesel type were being installed. Some were used to produce electricity for small municipalities, while others were installed in water pumping plants. Many provided power for tugs, fishing boats, trawlers, and workboats.

In the early 1920s the General Electric Company suggested to the Ingersoll–Rand Company, for whom Price was working, that they cooperate in the building of a diesel-electric locomotive. At that time many of the locomotives in service were powered by gasoline engines. A diesel-electric locomotive with Price's engine was completed in 1924 and placed in service for switching purposes in New York City. The success of this locomotive resulted in orders from railroads, factories, and open-pit mines. The engine used in most of these installations was a six-cylinder, 25-centimetre bore, 30-centimetre stroke system, rated 300 brake horsepower at 600 revolutions and weighing 6,800 kilograms.

Diesel-electric locomotive

Subsequent developments and applications. Many diesel engines were purchased for marine propulsion. The diesels, however, normally rotated faster than was desirable for the propellers of large ships because the high speeds of the huge propellers tended to create hollowed-out areas within the water around the propeller (cavitation), with resultant loss of thrust. The problem did not exist, however, with smaller propellers, and diesel engines proved especially suitable for yachts, in which speed is desired. The problem was solved by utilizing a diesel-electric installation in which the engines were connected to direct-current generators that furnished the electricity to drive an electric motor connected to the ship's propeller. There were also many installations in which the diesel was connected either directly or through gears to the propeller. When diesel engines of larger horsepower and slower rotation speeds became available, they were installed in cargo and passenger ships.

The diesel engine became the predominant power plant for military equipment on the ground and at sea during World War II. Since then it has been adopted for

use in heavy construction machinery, high-powered farm tractors, and most large trucks and buses. Diesel engines also have been installed in hospitals, telephone exchanges, airports, and various other facilities to provide emergency power during electrical power outages. In addition, they have been used in automobiles, albeit on a limited scale. Although diesels provide better fuel economy than gasoline engines, they do not run as smoothly as the latter and emit higher levels of pollutants. (L.V.A./C.L.P.II)

GAS-TURBINE ENGINES

General characteristics. Although the term gas turbine literally refers only to a turbine that employs a gas as the working fluid, it is conventionally used to describe a complete internal-combustion engine consisting of at least a compressor, a combustion chamber, and a turbine. Useful work or propulsive thrust can be obtained from the engine. It may drive a generator, pump, or propeller or, in the case of a pure jet aircraft engine, develop thrust by accelerating the turbine exhaust flow through a nozzle. Large amounts of power can be produced by a gas-turbine engine which, for the same output, is much smaller and lighter than a reciprocating internal-combustion engine. Reciprocating engines depend on the up-and-down motion of a piston, which must then be converted to rotary motion by a crankshaft arrangement, whereas a gas turbine delivers rotary shaft power directly. Although conceptually the gas-turbine engine is a simple device, the components for an efficient unit must be carefully designed and manufactured from costly materials because of the high temperatures and stresses encountered during operation. Thus, gas-turbine engine installations are usually limited to large units where they become cost-effective.

Differences between gas-turbine and reciprocating engines

Gas-turbine engine cycles. *Idealized simple open-cycle gas-turbine engine.* Most gas turbines operate on an open cycle in which air is taken from the atmosphere, compressed in a centrifugal or axial-flow compressor, and then fed into a combustion chamber. Here, fuel is added and burned at an essentially constant pressure with a portion of the air. Additional compressed air, which is bypassed around the burning section and then mixed with the very hot combustion gases, is required to keep the combustion chamber exit (in effect, the turbine inlet) temperature low enough to allow the turbine to operate continuously. If the unit is to produce shaft power, the combustion products (mostly air) are expanded in the turbine to atmospheric pressure. Most of the turbine output is required to operate the compressor; only the remainder is available to supply shaft work to a generator, pump, or other device. In a jet engine the turbine is designed to provide just enough output to drive the compressor and auxiliary devices. The stream of gas then leaves the turbine at an intermediate pressure (above local atmospheric pressure) and is fed through a nozzle to produce thrust. A simplified schematic for a gas turbine engine is given in Figure 28. Pressure–volume relations are also shown in the diagram.

By courtesy of Solar Division International Harvester Company

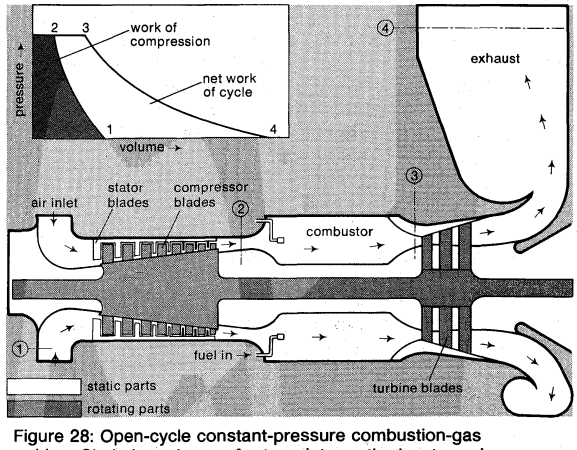


Figure 28: Open-cycle constant-pressure combustion-gas turbine. Circled numbers refer to points on the inset graph of the pressure-volume relationship during a working cycle (see text).

An idealized gas-turbine engine operating without any losses on this simple Brayton cycle is considered first. If, for example, air enters the compressor at 15° C and atmospheric pressure and is compressed to one megapascal, it then absorbs heat from the fuel at a constant pressure until the temperature reaches 1,100° C prior to expansion through the turbine back to atmospheric pressure. This idealized unit would require a turbine output of 1.68 kilowatts for each kilowatt of useful power with 0.68 kilowatt absorbed to drive the compressor. The thermal efficiency of the unit (net work produced divided by energy added through the fuel) would be 48 percent.

Brayton cycle

Actual simple open-cycle performance. If for a unit operating between the same pressure and temperature limits the compressor and the turbine are only 80 percent efficient (*i.e.,* the work of an ideal compressor equals 0.8 times the actual work, while the actual turbine output is 0.8 times the ideal output), the situation changes drastically even if all other components remain ideal. For every kilowatt of net power produced, the turbine must now produce 2.71 kilowatts while the compressor work becomes 1.71 kilowatts. The thermal efficiency drops to 25.9 percent. This illustrates the importance of highly efficient compressors and turbines. Historically it was the difficulty of designing efficient compressors, even more than efficient turbines, that delayed the development of the gas-turbine engine. Modern units can have compressor efficiencies of 86–88 percent and turbine efficiencies of 88–90 percent at design conditions.

Efficiency and power output can be increased by raising the turbine-inlet temperature. All materials lose strength at very high temperatures, however, and since turbine blades travel at high speeds and are subject to severe centrifugal stresses, turbine-inlet temperatures above 1,100° C require special blade cooling. It can be shown that for every maximum turbine-inlet temperature there is also an optimum pressure ratio. Modern aircraft gas turbines with blade cooling operate at turbine-inlet temperatures above 1,370° C and at pressure ratios of about 30:1.

Increasing efficiency and power output

Intercooling, reheating, and regeneration. In aircraft gas-turbine engines attention must be paid to weight and diameter size. This does not permit the addition of more equipment to improve performance. Accordingly, commercial aircraft engines operate on the simple Brayton cycle idealized above. These limitations do not apply to stationary gas turbines where components may be added to increase efficiency. Improvements could include (1) decreasing compression work by intermediate cooling, (2) increasing turbine output by reheating after partial expansion, or (3) decreasing fuel consumption by regeneration.

The first improvement would involve compressing air at nearly constant temperature. Although this cannot be achieved in practice, it can be approximated by intercooling (*i.e.,* by compressing the air in two or more steps and water-cooling it between steps back to its initial temperature). Cooling decreases the volume of air to be handled and, with it, the compression work required.

The second improvement involves reheating the air after partial expansion through a high-pressure turbine in a second set of combustion chambers before feeding it into a low-pressure turbine for final expansion. This process is similar to the reheating used in a steam turbine.

Both approaches require considerable additional equipment and are used less frequently than the third improvement. Here, the hot exhaust gases from the turbine are passed through a heat exchanger, or regenerator, to increase the temperature of the air leaving the compressor prior to combustion. This reduces the amount of fuel needed to reach the desired turbine-inlet temperature. The increase in efficiency is, however, tied to a large increase in initial cost and will be economical only for units that are run almost continuously.

Use of regenerators

Major components of gas-turbine engines. *Compressor.* Early gas turbines employed centrifugal compressors, which are relatively simple and inexpensive. They are, however, limited to low pressure ratios and cannot match the efficiencies of modern axial-flow compressors. Accordingly, centrifugal compressors are used today primarily in small industrial units.

An axial-flow compressor is the reverse of a reaction turbine (see *Steam turbines* above). The blade passages, which look like twisted, highly curved airfoils, must exert a tangential force on the fluid with the pressures on one side of the blade higher than on the other. For subsonic flow, an increase in pressure requires the flow area to also increase, thus reducing the flow velocity between the blade passages and diffusing the flow. A typical compressor stage is shown schematically in Figure 29 with corresponding velocity diagrams. A simple passage flow interpretation, however, is not enough for design purposes. Here, a row of compressor blades must be viewed as a set of closely spaced, highly curved airfoil shapes with which airflow strongly interacts. There will not only be a rise in pressure along the blades but a variation between them as well. Flow friction, leakage, wakes produced by the previous sets of blades, and secondary circulation or swirl flows all contribute to losses in a real unit. Tests of stationary blade assemblies, known as cascades, can be performed in special wind tunnels, but actual blade arrangements in a rotating assembly require special test setups or rigs.

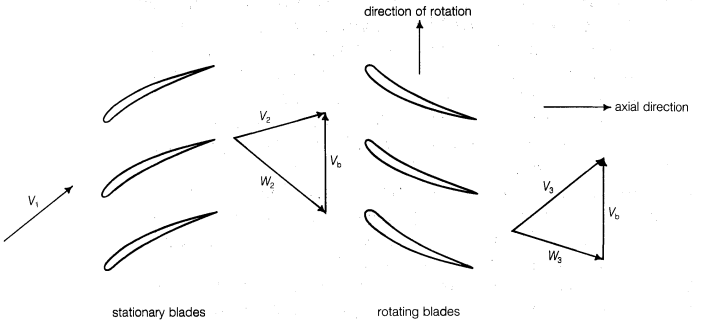


Figure 29: *Typical axial-flow compressor stage with velocity diagrams.*
Here, V is the absolute velocity of fluid, V_b is the blade velocity, and W is the velocity of fluid relative to the blade. Subscript 1 denotes entry into stationary blade (stator), subscript 2 signifies exit from stator or entry into rotor, and subscript 3 indicates exit from rotor.

Blades must be designed not only to have the correct aerodynamic shape but also to be light and not prone to critical vibrations. Recent advances in compressor (and turbine) blade design have been aided by extensive computer programs.

While moderately large expansion-pressure ratios can be achieved in a reaction-turbine stage, only relatively small pressure increases can be handled by a compressor stage—typically pressure ratios per stage of 1.35 or 1.4 to 1 in a modern design. Thus, compressors require more stages than turbines. If higher stage pressure ratios are attempted, the flow will tend to separate from the blades, leading to turbulence, reduced pressure rise, and a "stalling" of the compressor with a concurrent loss of engine power. Unfortunately, compressors are most efficient close to this so-called surge condition, where small disturbances can disrupt operation. It remains a major challenge to the designer to maintain high efficiency without stalling the compressor.

As the air is compressed, its volume decreases. Thus the annular passage area should also decrease if the through-flow velocity is to be kept nearly constant—*i.e.*, the blades have to become shorter at higher pressures. An optimum balance of blade-tip speeds and airflow velocities often requires that the rotational speed of the front, low-pressure end of the compressor be less than that of the high-pressure end. This is achieved in large aircraft gas turbines by "spooled" shafts where the shaft for the low-pressure end, driven by the low-pressure portion of the turbine, is running at a different speed within the hollow high-pressure compressor/turbine shaft, with each shaft having its own bearings. Both twin- and triple-spool engines have been developed.

Combustion chamber. Air leaving the compressor must first be slowed down and then split into two streams. The smaller stream is fed centrally into a region where atomized fuel is injected and burned with a flame held in place by a turbulence-generating obstruction. The larger, cooler stream is then fed into the chamber through holes along a "combustion liner" (a sort of shell) to reduce the overall temperature to a level suitable for the turbine inlet. Combustion can be carried out in a series of nearly cylindrical elements spaced around the circumference of the engine called cans, or in a single annular passage with fuel-injection nozzles at various circumferential positions. The difficulty of achieving nearly uniform exit-temperature distributions in a short aircraft combustion chamber can be alleviated in stationary applications by longer chambers with partial internal reversed flow.

Turbine. The turbine is normally based on the reaction principle with the hot gases expanding through up to eight stages using one- or two-spooled turbines. In a turbine driving an external load, part of the expansion frequently takes place in a high-pressure turbine that drives only the compressor while the remaining expansion takes place in a separate, "free" turbine connected to the load.

One- or two-spooled turbines

High-performance aircraft engines usually employ multiple spools. A recent large aircraft-engine design operating with an overall pressure ratio of 30.5:1 uses two high-turbine pressure stages to drive 11 high-pressure compressor stages on the outer spool, rotating at 9,860 revolutions per minute, while four low-pressure turbine stages drive the fan for the bypass air as well as four additional low-pressure compressor stages through the inner spool turning at 3,600 revolutions per minute (see below). For stationary units, a total of three to five total turbine stages is more typical.

High temperatures at the turbine inlet and high centrifugal blade stresses necessitate the use of special metallic alloys for the turbine blades. (Such alloys are sometimes grown as single crystals.) Blades subject to very high temperatures also must be cooled by colder air drawn directly from the compressor and fed through internal passages. Two processes are currently used: (1) jet impingement on the inside of hollow blades, and (2) bleeding of air through tiny holes to form a cooling blanket over the outside of the blades.

Control and start-up. In a gas-turbine engine driving an electric generator, the speed must be kept constant regardless of the electrical load. A decrease in load from the design maximum can be matched by burning less fuel while keeping the engine speed constant. Fuel flow reduction will lower the exit temperature of the combustion chamber and, with it, the enthalpy drop available to the turbine. Although this reduces the turbine efficiency slightly, it does not affect the compressor, which still handles the same amount of air. The foregoing method of control is substantially different from that of a steam turbine, where the mass flow rate has to be changed to match varying loads.

An aircraft gas-turbine engine is more difficult to control. The required thrust, and with it engine speed, may have to be changed as altitude and aircraft speed are altered. Higher altitudes lead to lower air-inlet temperatures and pressures and reduce the mass flow rate through the engine. Aircraft now use complex computer-driven controls to adjust engine speed and fuel flow while all critical conditions are monitored continuously.

For start-up, gas turbines require an external motor which may either be electric or, for stationary applications, a small diesel engine.

External motor for start-up

Other design considerations. Many other aspects enter into the design of a modern gas-turbine engine, of which only a few examples can be given. Much attention must be paid, especially in a multispool unit, to the design of all bearings, including the thrust bearings that absorb axial forces, and to the lubrication system. As an engine is started up and becomes hot, components elongate or "grow," thereby affecting passage clearances and seals. Other considerations include bleeding air from the compressor and ducting it for turbine-blade cooling or for driving accessories.

Applications. By far the most important use of gas turbines is in aviation, where they provide the motive power for jet propulsion. Because of the significance of this application and the diversity of modern jet engines, the

subject will be dealt with at length in a separate section of the article. The present discussion will touch on the use of gas turbines in electric power generation and in certain industrial processes, as well as consider their role in marine, locomotive, and automotive propulsion.

Electric power generation. In the field of electric power generation, gas turbines must compete with steam turbines in large central power stations and with diesel engines in smaller plants. Even though the initial cost of a gas turbine is less than either alternative for moderately sized units, its inherent efficiency is also lower. Yet, a gas-turbine unit requires less space, and it can be placed on-line within minutes, as opposed to a steam unit that requires many hours for start-up. As a consequence, gas-turbine engines have been widely used as medium-sized "peak load" plants to run intermittently during short durations of high power demand on an electric system. In this case, initial costs, rather than fuel charges, become the prime consideration.

Advantages of a gas turbine

Early commercial stationary plants employed aircraft units operating at reduced turbine-inlet temperatures. The high rotational speed of aircraft turbines required special gearing to drive electric generators. More recently, special units have been designed for direct operation (in the United States) at 3,600 revolutions per minute. Units in sizes up to 200,000 kilowatts have been built, although the majority of installations are less than 100,000 kilowatts. These turbines have operated up to 6,000 hours per year on either liquid fuels or natural gas. Typical turbine-inlet temperatures for large units range from about 980° to 1,260° C with turbine blade cooling used at the higher temperatures.

Efficiency can be improved by adding a regenerator to exploit the high turbine exhaust temperatures (typically about 480° to 590° C). Alternatively, if the gas turbine serves as a peak-load unit for a continuously running steam power plant, the hot exhaust gases can be used to preheat by means of a heat exchanger the combustion air entering a steam boiler. A modern development involves feeding the gas turbine exhaust directly into a steam generator where additional fuel is burned, producing steam of moderate pressure for a steam turbine. An overall thermal efficiency of nearly 50 percent is claimed for these combined units, making them the most fuel-efficient power plants currently available.

Use as a peak-load unit for a steam power plant

Industrial uses. With sizes typically ranging from 1,000 to 50,000 horsepower, industrial gas-turbine engines can be used for many applications. These include driving compressors for pumping natural gas through pipelines, where a small part of the pumped gas serves as the fuel. Such units can be automated so that only occasional on-site supervision is required. A gas turbine can also be incorporated in an oil refining process called the Houdry process, in which pressurized air is passed over a catalyst to burn off accumulated carbon. The hot gases then drive a turbine directly without a combustion chamber. The turbine, in turn, drives a compressor to pressurize the air for the process. Small portable gas turbines with centrifugal compressors also have been used to operate pumps.

Marine propulsion. In this area of application, the gas-turbine engine has two advantages over steam- and diesel-driven plants: it is lightweight and compact. During the early 1970s a ship powered by a gas turbine capable of 20,000 horsepower was successfully tested at sea by the U.S. Navy over a period of more than 5,000 hours. Gas turbines were subsequently selected to power various new U.S. naval vessels.

Use in naval vessels

Locomotive propulsion. During the 1950s and '60s, manufacturers of locomotives built a number of vehicles powered by gas-turbine engines that use heavy oil. Although gas-turbine locomotives have had moderate success for long sustained runs, they have not been able to make significant inroads against diesel locomotives under normal running conditions, especially after increases in the relative cost of heavy fuel oils. Moreover, the inherent low efficiency of a simple open-cycle gas turbine becomes even worse at part-load or during idling when considerable fuel is needed to drive the compressor while producing little or no useful power.

Automotive propulsion. Gas-turbine engines were proposed for use in automobiles from the early 1960s. In spite of their small size and weight for a given power output and their low exhaust emissions compared to gasoline engines, the disadvantages of high manufacturing costs, low thermal efficiency, and poor part-load and idling performance have proven gas-turbine cars to be uneconomical and impractical.

Development of gas turbines. *Origins.* The earliest device for extracting rotary mechanical energy from a flowing gas stream was the windmill (see above). It was followed by the smokejack, first sketched by Leonardo da Vinci and subsequently described in detail by John Wilkins, an English clergyman, in 1648. This device consisted of a number of horizontal sails that were mounted on a vertical shaft and driven by the hot air rising from a chimney. With the aid of a simple gearing system, the smokejack was used to turn a roasting spit.

The smokejack

Various impulse and reaction air-turbine drives were developed during the 19th century. These made use of air, compressed externally by a reciprocating compressor, to drive rotary drills, saws, and other devices. Many such units are still being used, but they have little in common with the modern gas-turbine engine, which includes a compressor, combustion chamber, and turbine to make up a self-contained prime mover. The first patent to approximate such a system was issued to John Barber of England in 1791. Barber's design called for separate reciprocating compressors whose output air was directed through a fuel-fired combustion chamber. The hot jet was then played through nozzles onto an impulse wheel. The power produced was to be sufficient to drive both the compressor and an external load. No working model was ever built, but Barber's sketches and the low efficiency of the components available at the time make it clear that the device could not have worked even though it incorporated the essential components of today's gas-turbine engine.

Although many devices were subsequently proposed, the first significant advance was covered in an 1872 patent granted to F. Stolze of Germany. Dubbed the fire turbine, his machine consisted of a multistage, axial-flow air compressor that was mounted on the same shaft as a multistage, reaction turbine. Air from the compressor passed through a heat exchanger, where it was heated by the turbine exhaust gases before passing through a separately fired combustion chamber. The hot compressed air was then ducted to the turbine. Although Stolze's device anticipated almost every feature of a modern gas-turbine engine, both compressor and turbine lacked the necessary efficiencies to sustain operation at the limited turbine-inlet temperature possible at the time.

Developments of the early 20th century. The first successful gas turbine, built in Paris in 1903, consisted of a three-cylinder, multistage reciprocating compressor, a combustion chamber, and an impulse turbine. It operated in the following way: Air supplied by the compressor was burned in the combustion chamber with liquid fuel. The resulting gases were cooled somewhat by the injection of water and then fed to an impulse turbine. This system, which had a thermal efficiency of about 3 percent, demonstrated for the first time the feasibility of a practical gas-turbine engine.

First successful gas turbine

Two other devices with intermittent gas action, both developed at about the same time, deserve mention. A 10,000-revolutions-per-minute unit built in Paris in 1908 had four explosion chambers located on the periphery of a de Laval impulse turbine. Each chamber, containing air and fuel, was fired sequentially to provide a nearly continuous flow of high-temperature, high-pressure gases that were fed through nozzles to the turbine wheel. The momentary partial vacuum created by the hot gases rushing from the explosion chamber was used to draw in a new charge of air.

Of greater significance was the "explosion" turbine developed by Hans Holzwarth of Germany, whose initial experiments started in 1905. In this system, a compressor introduced a charge of air and fuel into a constant-volume combustion chamber. After ignition, the hot, high-pressure gas escaped through spring-loaded valves into nozzles directed against the blading of a turbine. The valves re-

mained open until the gas was discharged, at which point a fresh charge was brought into the combustion chamber. Since the pressure increase in the compressor was only about one-fourth of the maximum pressure reached after combustion, the unit could operate even though the compressor efficiency was low. Holzwarth and various collaborators continued to develop the explosion turbine for more than 30 years until it was eventually superseded by the modern gas-turbine engine.

System featuring both a rotating compressor and a turbine

To be successful, a steady-flow engine based on the ideas first proposed by Stolze depends not only on high efficiencies (more than 80 percent) for both the rotating compressor and the turbine but also on moderately high turbine-inlet temperatures. The first successful experimental gas turbine using both rotary compressors and turbines was built in 1903 by Aegidus Elling of Norway. In this machine, part of the air leaving a centrifugal compressor was bled off for external power use. The remainder, which was required to drive the turbine, passed through a combustion chamber and then through a steam generator where the hot gas was partially cooled. This combustion gas was cooled further (by steam injected into it) to 400° C, the maximum temperature that Elling's radial-inflow turbine could handle. The earliest operational turbine of this type delivered 11 horsepower. Many subsequent improvements led to another experimental Elling turbine, which by 1932 could produce 75 horsepower. It employed a compressor with 71-percent efficiency and a turbine with an efficiency of 82 percent operating at an inlet temperature of 550° C. Norway's industry, however, was unable to capitalize on these developments, and no commercial units were built. The first industrial success did not come until 1936, when the Swiss firm of Brown Boveri independently developed a gas turbine for the Houdry process (see above).

Also during the mid-1930s a group headed by Frank Whittle at the British Royal Aircraft Establishment (RAE) undertook efforts to design an efficient gas turbine for jet propulsion of aircraft. The unit produced by Whittle's group worked successfully during tests; it was determined that a pressure ratio of about 4 could be realized with a single centrifugal compressor running at roughly 17,000 revolutions per minute. Shortly after Whittle's achievement, another RAE group, led by A.A. Griffith and H. Constant, began developmental work on an axial-flow compressor. Axial-flow compressors, though much more complex and costly, were better suited for detailed blade-design analysis and could reach higher pressures and flow rates and, eventually, higher efficiencies than their centrifugal counterparts.

Independent parallel developments in Germany, initiated by Hans P. von Ohain working with the manufacturing firm of Ernst Heinkel, resulted in a fully operational jet aircraft engine that featured a single centrifugal compressor and a radial-inflow turbine. This engine was successfully tested in the world's first jet-powered airplane flight on Aug. 27, 1939. Subsequent German developments directed by Anselm Franz led to the Junkers Jumo 004 engine for the Messerschmitt Me-262 aircraft, which was first flown in 1942. In Germany as well as in Britain, the search for higher temperature materials and longer engine life was aided by experience gained in developing aircraft turbosuperchargers.

Use in military aircraft

Before the end of World War II gas-turbine jet engines built by Britain, Germany, and the United States were flown in combat aircraft. Within the next few decades both propeller-driven gas-turbine engines (turboprops) and pure jet engines developed rapidly, with the latter assuming an ever larger role as airplane speeds increased.

Recent trends. Because of the significant advances in gas-turbine engine design in the years following World War II, it was expected that such systems would become an important prime mover in many areas of application. However, the high cost of efficient compressors and turbines, coupled with the continued need for moderate turbine-inlet temperatures, have limited the adoption of gas-turbine engines. Their preeminence remains assured only in the field of aircraft propulsion for medium and large planes that operate at either subsonic or supersonic speeds. As for electric power generation, large central power plants that use steam or hydraulic turbines are expected to continue to predominate. The prospects appear bright, nonetheless, for medium-sized plants employing gas-turbine engines in combination with steam turbines. Further use of gas-turbine engines for peak power production is likely as well. These turbine engines also remain attractive for small and medium-sized, high-speed marine vessels and for certain industrial applications. (Fr.L.)

JET ENGINES

General characteristics. The prime mover of virtually all jet engines is a gas turbine. Variously called the core, gas producer, gasifier, or gas generator, the gas turbine converts the energy derived from the combustion of a liquid hydrocarbon fuel to mechanical energy in the form of a high-pressure, high-temperature airstream. This energy is then harnessed by what is termed the propulsor (*e.g.*, airplane propeller and helicopter rotor) to generate a thrust with which to propel the aircraft.

Principles of operation. *The prime mover.* The gas turbine operates on the Brayton cycle in which the working fluid is a continuous flow of air ingested into the engine's inlet (see *Gas-turbine engine cycles* above). As shown in Figure 30, the air is first compressed by a turbocompressor to a pressure ratio of typically 10 to 40 times the pressure of the inlet airstream. It then flows into a combustion chamber, where a steady stream of the hydrocarbon fuel, in the form of liquid spray droplets and vapour or both,

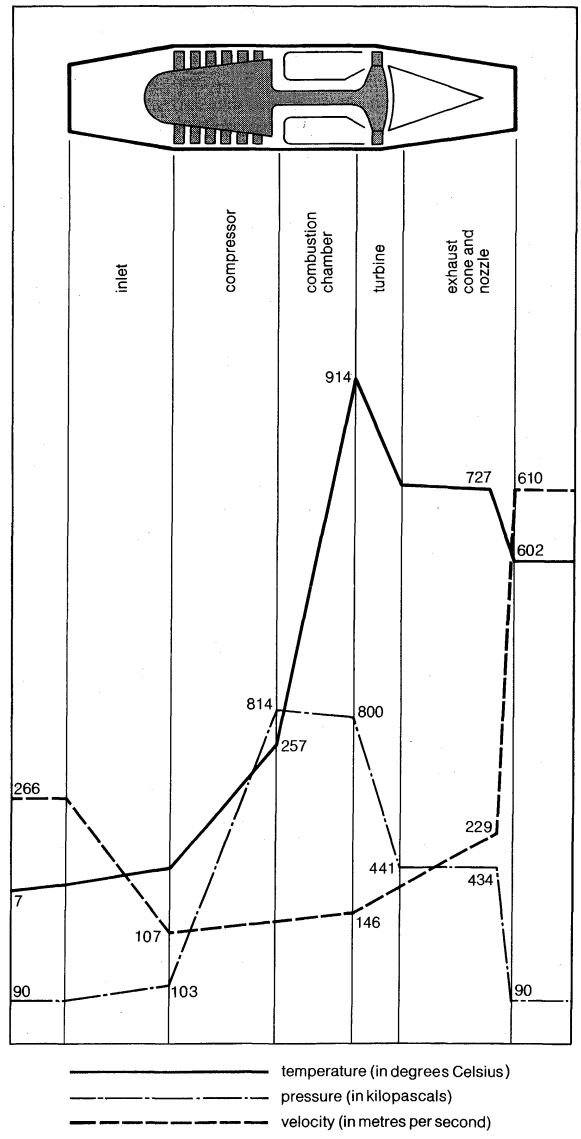


Figure 30: Cross section of a turbojet and (below) graph of typical operating conditions for its working fluid.

is introduced and burned at approximately constant pressure. This gives rise to a continuous stream of high-pressure combustion products whose average temperature is typically from 980° to 1,540° C or higher. This stream of gases flows through a turbine, which is linked by a torque shaft to the compressor and which extracts energy from the gas stream to drive the compressor. Because heat has been added to the working fluid at high pressure, the gas stream that exits from the gas generator after having been expanded through the turbine contains a considerable amount of surplus energy—*i.e.*, gas horsepower—by virtue of its high pressure, high temperature, and high velocity, which may be harnessed for propulsion purposes.

Gas horsepower

The heat released by burning a typical jet fuel in air is approximately 43,370 kilojoules per kilogram (18,650 British thermal units per pound) of fuel. If this process were 100 percent efficient, it would then produce a gas power for every unit of fuel flow of 7.45 horsepower/(pounds per hour), or 12 kilowatts/(kilograms per hour). In actual fact, certain practical thermodynamic limitations, which are a function of the peak gas temperature achieved in the cycle, restrict the efficiency of the process to about 40 percent of this ideal value. The peak pressure achieved in the cycle also affects the efficiency of energy generation. This implies that the lower limit of specific fuel consumption (SFC) for an engine producing gas horsepower is 0.336 (pound per hour)/horsepower, or 0.207 (kilogram per hour)/kilowatt. In actual practice, the SFC is even higher than this lower limit because of inefficiencies, losses, and leakages in the individual components of the prime mover.

Because weight and volume are at a premium in the overall design of an aircraft and because the power plant represents a large fraction of any aircraft's total weight and volume, these parameters must be minimized in the engine design. The airflow that passes through an engine is a representative measure of the engine's cross-sectional area and hence its weight and volume. Therefore, an important figure of merit for the prime mover is its specific power—the amount of power that it generates per unit of airflow. This quantity is a very strong function of the peak gas temperature in the core at the discharge of the combustion chamber. Modern engines generate from 150 to 250 horsepower/(pound per second), or 247 to 411 kilowatts/(kilogram per second).

Specific power

The propulsor. The gas horsepower generated by the prime mover in the form of hot, high-pressure gas is used to drive the propulsor, enabling it to generate thrust for propelling or lifting the aircraft. The principle on which such a thrust is produced is based on Newton's second law of motion. This law generalizes the observation that the force (F) required to accelerate a discrete mass (m) is proportional to the product of that mass and the acceleration (a). In effect,

$$F = ma = \frac{wa}{g},$$

where the mass is taken as the weight (w) of the object divided by the acceleration due to gravity (g) at the place where the object was weighed. In the case of a jet engine, one is generally dealing with the acceleration of a steady stream of air rather than with a discrete mass. Here, the equivalent statement of the second law of motion is that the force (F) required to increase the velocity of a stream of fluid is proportional to the product of the rate of mass flow (M) of the stream and the change in velocity of the stream,

$$F = M(V_j - V_0) = \frac{W(V_j - V_0)}{g},$$

where the inlet velocity (V_0) relative to the engine is taken to be the flight velocity and the discharge velocity (V_j) is the exhaust or jet velocity relative to the engine. W is the rate of weight flow of working fluid (*i.e.*, air or products of combustion) divided by the acceleration of gravity in the place where the weight flow is measured. The relatively small effect of the weight flow of fuel in creating a difference between the weight flow of the inlet and exhaust streams is intentionally disregarded.

One thereby infers that the components of a propulsor must exert a force F on the stream of air flowing through the propulsor if this device accelerates the airstream from the flight velocity V_0 to the discharge velocity V_j. The reaction to that force F is ultimately transmitted by the mounts of the propulsor to the aircraft as propulsive thrust.

There are two general approaches to converting gas horsepower to propulsive thrust. In one, a second turbine (*i.e.*, a low-pressure, or power, turbine) may be introduced into the engine flow path to extract additional mechanical power from the available gas horsepower. This mechanical power may then be used to drive an external propulsor, such as an airplane propeller or helicopter rotor. In this case, the thrust is developed in the propulsor as it energizes and accelerates the airflow through the propulsor—*i.e.*, an airstream separate from that flowing through the prime mover.

Use of a second turbine

In the second approach, the high-energy stream delivered by the prime mover may be fed directly to a jet nozzle, which accelerates the gas stream to a very high velocity as it leaves the engine, as is typified by the turbojet. In this case, the thrust is developed in the components of the prime mover as they energize the gas stream.

In other types of engines, such as the turbofan, thrust is generated by both approaches: A major part of the thrust is derived from the fan, which is powered by a low-pressure turbine and which energizes and accelerates the bypass stream (see below). The remaining part of the total thrust is derived from the core stream, which is exhausted through a jet nozzle.

Just as the prime mover is an imperfect device for converting the heat of fuel combustion to gas horsepower, so the propulsor is an imperfect device for converting the gas horsepower to propulsive thrust. There is generally a great deal of energy left in the high-temperature, high-velocity jet stream exiting from the propulsor that is not fully exploited for propulsion. The efficiency of a propulsor, propulsive efficiency η_p, is the portion of the available energy that is usefully applied in propelling the aircraft compared to the total energy of the jet stream. For the simple but representative case of the discharge airflow equal to the inlet gas flow, it is found that

Propulsive efficiency

$$\eta_p = \frac{2V_0}{V_j + V_0}.$$

Although the jet velocity V_j must be larger than the aircraft velocity V_0 to generate useful thrust, a large jet velocity that exceeds flight speed by a substantial margin can be very detrimental to propulsive efficiency. Maximum propulsive efficiency is approached when the jet velocity is almost equal to (but, of necessity, slightly higher than) the flight speed. This fundamental fact has given rise to a large variety of jet engines, each designed to generate a specific range of jet velocities that matches the range of flight speeds of the aircraft that it is supposed to power (see Figure 31).

The net assessment of the efficiency of a jet engine is the measurement of its rate of fuel consumption per unit of thrust generated (*e.g.*, in terms of pounds, or kilograms, per hour of fuel consumed per pounds, or kilograms, of thrust generated). There is no simple generalization of the value of specific fuel consumption of a thrust engine. It is

Specific fuel consumption

Adapted from *Aerospace America* (October 1984)

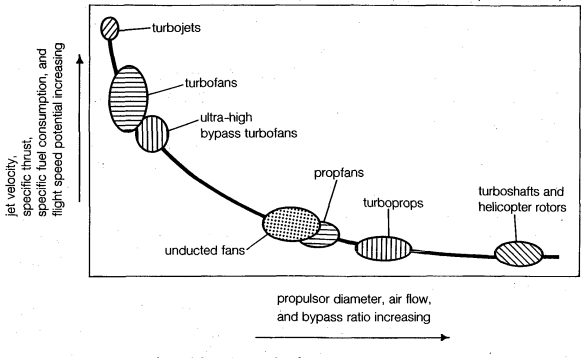


Figure 31: Effect of varying jet velocity on engine design and operation.

not only a strong function of the prime mover's efficiency (and hence its pressure ratio and peak-cycle temperature) but also of the propulsive efficiency of the propulsor (and hence of the engine type). It also is a strong function of the aircraft flight speed and the ambient temperature (which is in turn a strong function of altitude, season, and latitude).

Basic engine types. Achieving a high propulsive efficiency for a jet engine is dependent on designing it so that the exiting jet velocity is not greatly in excess of the flight speed. At the same time, the amount of thrust generated is proportional to that very same velocity excess that must be minimized. This set of restrictive requirements has led to the evolution of a large number of specialized variations of the basic turbojet engine, each tailored to achieve a balance of good fuel efficiency, low weight, and compact size for duty in some band of the flight speed–altitude–mission spectrum. There are two major general features characteristic of all the different engine types, however. First, in order to achieve a high propulsive efficiency, the jet velocity, or the velocity of the gas stream exiting the propulsor, is matched to the flight speed of the aircraft—slow aircraft have engines with low jet velocities and fast aircraft have engines with high jet velocities. Second, as a result of designing the jet velocity to match the flight speed, the size of the propulsor varies inversely with the flight speed of the aircraft—slow aircraft have very large propulsors, as, for example, the helicopter rotor—and the relative size of the propulsor decreases with increasing design flight speed—turboprop propellers are relatively small and turbofan fans even smaller.

Although the turbojet is the simplest jet engine and was invented and flown first among all the engine types, it seems useful to examine the entire spectrum of engines in the order of the flight-speed band in which they serve, starting with the slowest—namely, the turboshaft engine, which powers helicopters.

Turboshaft engines. The helicopter is designed to operate for substantial periods of time hovering at zero flight speed. Even in forward flight, helicopters rarely exceed 240 kilometres per hour or a Mach number of 0.22. (The Mach number is the ratio of the velocity of the aircraft to the speed of sound.) The principal propulsor is the helicopter rotor, which is driven by one or more turboshaft engines (Figure 32) in all modern helicopters of large size. As was previously noted, the propulsor is designed to give a very low discharge or jet velocity and is by the same token very large for a given size aircraft when compared to the propulsors of higher-speed aircraft. The prime mover of a helicopter is a core engine whose gas horsepower is extracted by a power turbine, which then drives the helicopter rotor via a speed-reducing (and combining) gearbox. The power turbine is usually located on a spool separate from the gas generator; thus its rotative speed and that of the helicopter rotor which it drives are independent of the rotative speed of the gas generator. This allows the rotor speed to be varied or kept constant independently of the gas-generator speed, which must be varied to modulate the amount of power generated.

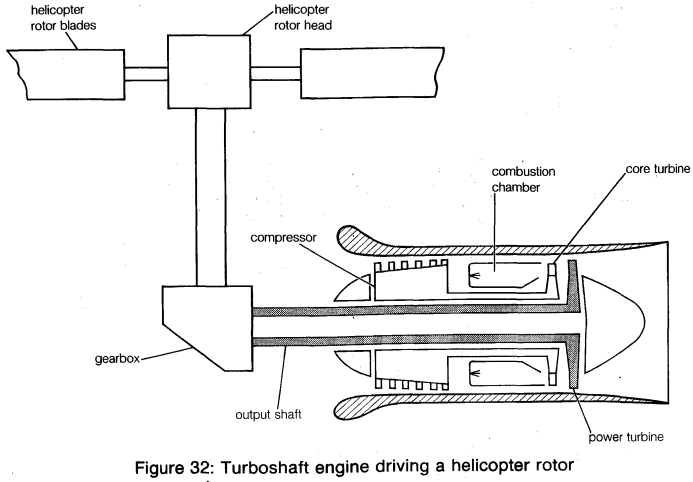


Figure 32: Turboshaft engine driving a helicopter rotor as propulsor.

Turboprops, propfans, and unducted fan engines. The turboprop is the power plant that occupies the next band of flight speeds in the flight spectrum, from a Mach number of 0.2 to 0.7. The propulsor is a propeller with a somewhat higher discharge, or jet velocity, than that of the helicopter rotor to match the flight speed, and it has a proportionately smaller area than the latter for a similarly sized aircraft. As shown in Figure 33, the prime mover is a turboshaft engine (very similar to the one that drives a helicopter rotor except for a different gearbox) designed to provide a somewhat higher rotative speed for the propeller, which turns faster than the helicopter rotor having a much larger diameter. The control mode of the turboprop also is somewhat different from that of a helicopter's turboshaft engine. In a helicopter the pilot calls for power by manipulating the pitch of the rotor blades (a greater pitch taking a bigger "bite" of air and so demanding more power to maintain rotative speed). The engine's control responds by increasing fuel to the engine to maintain output shaft speed. In a turboprop, the pilot calls for power by selecting fuel flow to the prime mover. The propeller control responds by varying propeller pitch to attain a greater "pull" while maintaining a preselected propeller rotative speed.

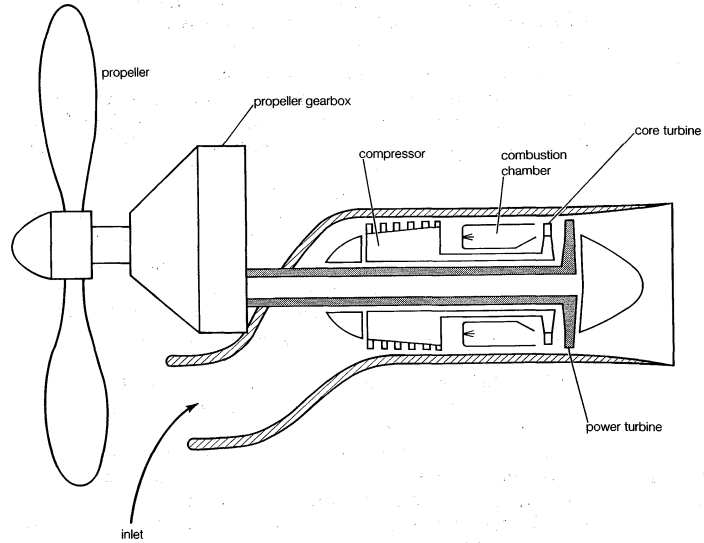


Figure 33: Turboprop engine driving a single rotation propeller as propulsor; tractor arrangement.

Trends in turboprop design

A recent trend in turboprop design has been the evolution of propellers for efficient operation at transonic flight speeds (those approximating the speed of sound), much higher than previously achieved—up to Mach numbers of 0.85. This usually involves a higher disk loading (*i.e.*, a higher discharge velocity from the propeller) in order to permit the use of a smaller diameter propeller. This trend has been accompanied by an increase in the number of blades in the propeller (from six to 12 instead of the more common two to four blades in lower-speed propellers). The blades are scimitar-shaped, with swept-back leading edges at the blade tips to accommodate the large Mach numbers encountered by the propeller tip at high rotative and flight speeds. Such high-speed propulsors are called propfans.

Another variation of the propulsor involves the application of two concentric propellers on the same centreline, driven by the same prime mover through a gearbox that causes each propeller to rotate in a direction opposite the other. Such counter-rotating propellers are capable of significantly higher propulsive efficiency and higher disk loading than conventional propellers.

Counter-rotating propellers

In most turboprop installations the prime mover is mounted on the wing, and the plane of the propeller is forward of the prime mover (the so-called tractor layout). Modern high-speed aircraft may find it more advantageous to mount the engine more toward the rear of the aircraft, with the plane of the propeller aft of the engine. These ar-

rangements are referred to as "pusher" layouts. A recently developed engine layout, identified as the unducted fan (or UDF; trademark), provides a set of very high-efficiency counter-rotating propeller blades, each blade mounted on one of either of two sets of counter-rotating low-pressure turbine stages and achieving all the advantages of the arrangement without the use of a gearbox (see Figure 34).

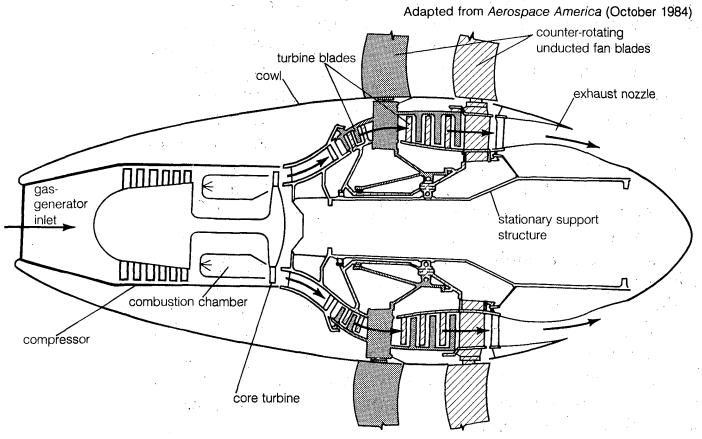


Figure 34: An unducted fan engine (UDF; trademark) with counter-rotating propellers, or unducted fan blades; pusher arrangement.

Medium-bypass turbofans, high-bypass turbofans, and ultrahigh-bypass engines. Moving up in the spectrum of flight speeds to the transonic regime—Mach numbers from 0.75 to 0.9—the most common engine configurations are turbofan engines, such as those in Figures 35A and 35B. In a turbofan, only a part of the gas horsepower generated by the core is extracted to drive a propulsor, which usually consists of a single low-pressure-ratio, shrouded turbocompression stage. The fan is generally placed in front of the core inlet so that the air entering the core first passes through the fan and is partially compressed by it. Most of the air, however, bypasses the core (hence the designation bypass stream) and goes directly to an exhaust nozzle. The core stream, with some modest fraction of the gas horsepower remaining (not extracted to drive the fan) proceeds directly to its own exhaust nozzle.

Bypass ratio

A key parameter for classifying the turbofan is its bypass ratio, defined as the ratio of the mass flow rate of the bypass stream to the mass flow rate entering the core. Since the highest propulsion efficiencies are obtained by the engines with the highest bypass ratios, one would expect to find all engines of that design in this flight speed regime. (Some of the variation derives from historical evolution.) In actuality, however, one finds engines with a broad spectrum of bypass ratios, including medium-bypass engines (with bypass ratios from 2 to 4), high-bypass engines (with bypass ratios from 5 to 8), and ultrahigh-bypass engines, so-called UBEs (with bypass ratios from 9 to 15 or higher). A whole generation of low- and medium-bypass engines has completely supplanted the first generation of aircraft powered by (zero-bypass) turbojet engines. Moreover, that generation was itself supplanted by a third generation of medium- and high-bypass turbofan engines. There are several other reasons why engines with less than the highest bypass ratios hypothetically achievable are still in use. Very high bypass ratios involve the use of fans with very large diameters, which in turn entail very heavy components; this increases the difficulty of installing the engine on aircraft and maintaining sufficient ground clearance. In addition, the weight and complexity of the apparatus required to reverse the direction of the bypass stream (to achieve thrust reversal in order to shorten the aircraft's landing roll) also increases with the bypass ratio. The long-term trend, however, is definitely toward higher and higher bypass ratios.

There are several unique features and ancillary devices found in turbofan engines. As shown in Figure 35A, ultrahigh-bypass engines may have a gearbox between the drive turbine and the fan to simplify the design of the small-diameter turbine (with the attendant high rotative speed) without compromising the performance of the very large-diameter fan (with the attendant low rotative speed). Variable-pitch fan blades are generally required for thrust reversal in such ultrahigh-bypass fans, while in medium- and high-bypass engines the thrust reversing is usually accomplished by introducing blocker doors into the bypass stream. In high- and medium-bypass turbofans such as is shown in Figure 35B, a small but significant improvement in propulsive efficiency can be achieved by mixing the airstream of the hot core and cold bypass streams before the total airstream enters a single jet nozzle.

Provisions for thrust reversing

Low-bypass turbofans and turbojets. In the next higher regime of aircraft flight speed, the low supersonic range from Mach numbers above 1 up to 2 or 3, one finds the application of the simple turbojet (with no bypass stream) and the low-bypass turbofan engine (with a bypass ratio up to 2), such as that pictured in Figure 36.

Although the low-bypass turbofan has the same general appearance as a turbofan with a larger bypass ratio, certain special features are unique to low-bypass engines. The lower total flow in the fan generally involves a higher fan pressure ratio (for equivalent amounts of energy available from the drive turbine), and so such a fan usually has more than one (*i.e.*, two or three) turbocompressor stages. Engines designed to operate at the low supersonic range generally have insufficient thrust in other flight regimes or modes where they must operate for short durations, as, for instance, acceleration through transonic speed, takeoff from high-altitude airports under conditions of extremely high temperatures and high gross weight, or combat maneuvers at high supersonic flight speed. Rather than installing a larger engine to meet these requirements, it is more effective to add an afterburner to a turbofan engine as a means of thrust augmentation (see Figure 36). The

Use of an afterburner

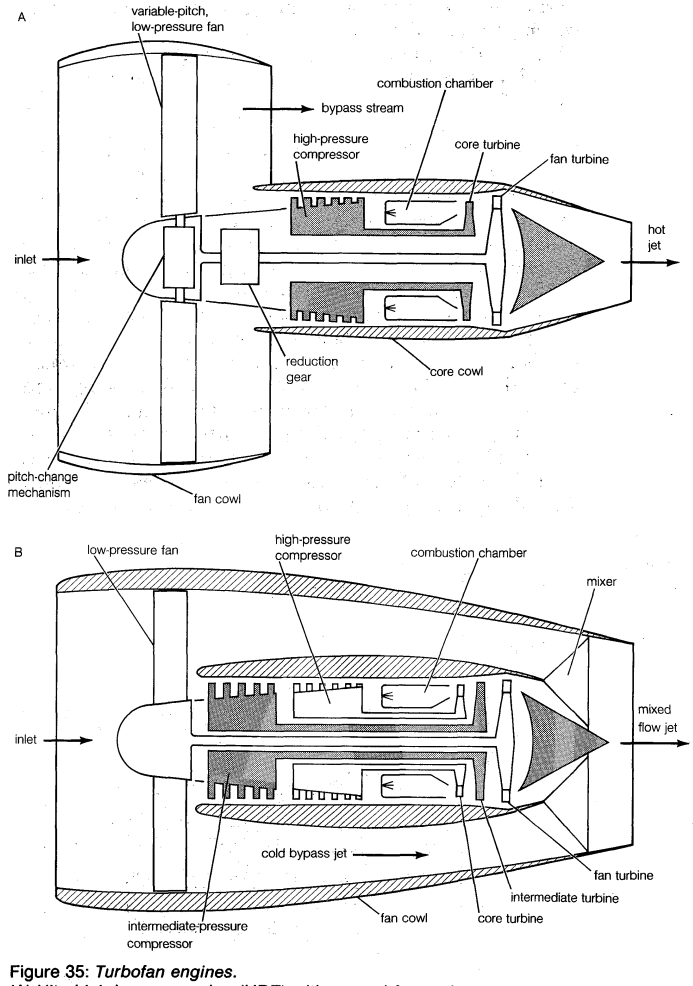


Figure 35: *Turbofan engines.*
(A) Ultrahigh-bypass engine (UBE) with geared fan and variable-pitch blading for thrust reversal. (B) High-bypass turbofan with two-spool core and mixed-flow jet.

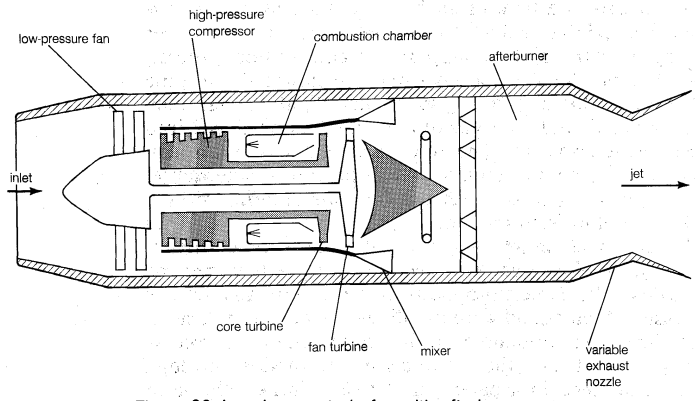


Figure 36: Low-bypass turbofan with afterburner.

afterburner is a secondary combustion system that operates in the exhaust stream of the engine before the stream is introduced into the exhaust nozzle. Such a device is not as fuel-efficient as the main turbofan section of the engine because heat addition occurs at a lower pressure than in the main burner. The afterburner, however, is relatively simple and lightweight, since it does not contain any rotating machinery. For the same reason, it may be operated to a much higher discharge temperature (typically 1,760° C), so that it is capable of augmenting the thrust of the turbofan by as much as 50 percent.

The afterburner in a turbofan usually requires a mixer for mixing the relatively cool bypass air with the hot core stream; the cooler air is otherwise difficult to burn in the low-pressure environment of an afterburner. Also, in both the turbojet and the turbofan with an afterburner, the exhaust nozzle must have a variable throat area to accommodate the large variations in volumetric flow rate between the very hot exhaust stream from the operating afterburner and the cooler airstream discharged from the engine when the afterburner is not in use. Engines intended for supersonic flight generally have a much lower compression-pressure ratio than higher-bypass machines intended for subsonic or transonic operation. A major contributor to this tendency is the additional pressure ratio developed in the engine's inlet as it slows down or diffuses the very high-speed airstream that is ingested as the engine's working fluid—the ram effect. At transonic flight speed this pressure ratio is almost 2:1, so that the engine's compressor may be built to provide that much less pressure where peak pressure is otherwise limiting.

Ram effect

Early generations of jet-propelled aircraft in this low supersonic flight regime were powered by turbojet engines, but subsequent generations built for the same flight regime have largely been equipped with low-bypass turbofans. This substitution of engine type was undertaken primarily because such aircraft expend a great deal of their fuel at subsonic flight speed (*e.g.*, in takeoff, climb, loiter, acceleration, approach, and landing), where the turbofan provides an advantage in propulsive efficiency.

Ramjets and supersonic combustion ramjets. As has been seen, ram pressure plays an increasingly important role in the thermodynamic cycle of power and thrust generation of the jet engine at supersonic flight speeds. For flight speeds above Mach 2.5 or 3, the ram-pressure ratio becomes so high that a turbocompressor is no longer necessary for efficient thrust generation. Indeed, the pressure ratio eventually rises to such high values that the associated high ram temperatures make it difficult or impossible to place high-speed rotating machinery in the flow path without prohibitive amounts of cooling provision. This combination of circumstances gives rise to the ramjet, a jet engine in which the pressure increase is attributable only to the ram effect of the high flight speed; no turbomachinery is involved, and the main thrust producer is an afterburner (see Figure 37).

Absence of a turbine

Ramjets are lightweight and simple power plants, making them ideal candidates for supersonic flight vehicles that are launched from other flight vehicles at extremely high speed. They are less suitable for use in vehicles that must be sufficiently self-powered for subsonic takeoff, climb, and acceleration to supersonic flight speed; the subsonic ram pressure is insufficient to produce any reasonable amount of thrust, and so alternative propulsion devices must be provided.

In the flight regime of Mach 4 or 5, it is usually efficient to decelerate the inlet airstream to subsonic velocity before it enters the combustion system. At still higher Mach numbers, such deceleration becomes more difficult and costly in terms of pressure losses, and it is necessary to make provision for the combustion chamber to burn its fuel in the supersonic airstream. Such specialized ramjets are called scramjets (for supersonic combustion ramjets) and are projected to be fueled by a cryogenically liquified gas (*e.g.*, hydrogen or methane) instead of a liquid hydrocarbon. The primary reason for doing so is to exploit the greater heat release per unit weight of fuels that have a higher ratio of hydrogen to carbon atoms than ordinary fractions of petroleum even though this gain is partly negated by the higher volume per unit of heat release of those same fuels. Another incentive for employing a very cold fuel is that it may be used as a heat sink for cooling a very high-speed (and hence very hot) engine and aircraft structure. The scramjet has an unusual feature: the inlet deceleration and exhaust acceleration occur largely outside the enclosed engine inlet and exhaust ducts against external aircraft surfaces in front of and to the rear of the engine. In effect, the engine itself is little more than a sophisticated supersonic combustion chamber.

Scramjets

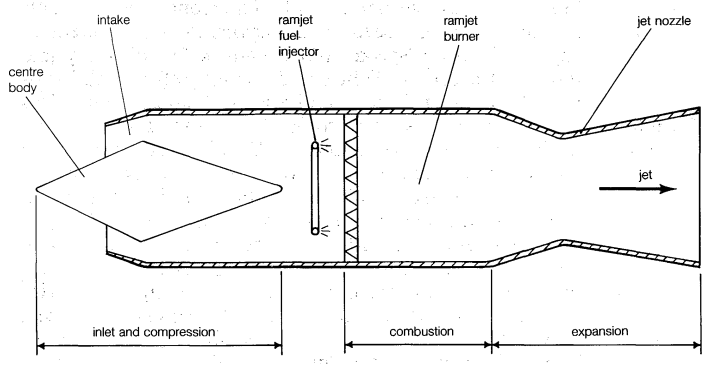


Figure 37: Arrangement of a ramjet.

Hybrid engine types. It is possible to tailor an engine configuration so that the engine is well suited for operation within a given band of the flight spectrum. To have an engine that will perform well in more than one band of the flight spectrum or in more than one regime of operation, it may be necessary to configure the power plant so that it can be converted from one engine type to another by means of variable geometry built into the engine components.

Vertical and short takeoff and landing (V/STOL) propulsion systems. Propulsion systems that provide aircraft with the capability of both vertical and conventional forward flight represent a formidable challenge to the engine designer. V/STOL aircraft have several major categories of engine arrangement. They are as follows:

Vertical and forward flight capability

1. As in a helicopter, the propulsor may consist of a rotor that is driven by one or more turboshaft engines and is installed in such a way as to provide vertical thrust. The entire aircraft must be tilted to give the thrust vector a forward component to achieve forward flight. This arrangement has certain limitations in terms of effectiveness, as borne out by the relative inefficiency of forward flight above a Mach number of 0.2.

2. The propulsors may be mounted on pivots so that they can be rotated from the position in which they give vertical thrust in a takeoff, hover, climb, descent, or landing maneuver and pivoted 90° to provide thrust for conventional forward flight (as in the tilt-rotor aircraft). The prime mover that drives the propulsor may either be tilted with the propulsor or be fixed in the wing and drive the tilting propulsor via a rotating shaft through the pivot axis. In some configurations, the entire wing of the aircraft, carrying fixed engines and propulsors, may be tilted as a single assembly.

3. The engines may be fixed in a position required to produce thrust for forward flight. Their exhaust systems, however, have built-in variable geometry, making it possible to vector the exhaust nozzle (or nozzles) or divert the exhaust gases by means of valves and auxiliary ducts to nozzles mounted in such a way as to provide vertical thrust or lift.

4. The aircraft may include two different sets of engines or propulsors (or both), fixed in position, with one set installed for forward flight and the other for vertical thrust (*i.e.*, the lift engines).

5. The aircraft may use a convertible engine. Such an engine has a single prime mover that is arranged to drive a fan for efficient forward propulsion, to drive a shaft that turns the main helicopter rotor, or to drive both a fan and a shaft. In order to convert from horizontal to vertical flight, variable-pitch fan blades or variable-pitch stators (or both) unload the fan, thereby making mechanical power available to drive the helicopter rotor for vertical movement.

Variable-cycle engines. For aircraft designed to fly mixed missions (*i.e.*, at subsonic, transonic, and supersonic flight speeds) with low levels of fuel consumption, it is desirable to have an engine with the characteristics of both a high-bypass engine (for subsonic flight speed) and a low-bypass engine (for supersonic flight speed). This requirement is typical for such high-speed commercial airliners as the Concorde, a type of supersonic transport built by the British and French. The Concorde is capable of traveling over oceans and unpopulated land areas at supersonic cruise speeds, but it cannot fly efficiently and quietly at subsonic flight speed for takeoff, ascent, cruising over populated areas, and approach and landing. This dual function is expected to be accomplished in the future by the variable-cycle engine (VCE). If the components of an engine are designed to accommodate the extreme limits of flow, pressure ratio, and other conditions involved in both high-bypass and low-bypass operation, the engine may be operated at either extreme of bypass ratio or at any bypass ratio between those extremes by means of a valve (or valves) in the bypass stream (in conjunction with a variable exhaust nozzle). When the valves are closed, they restrict the flow in the bypass stream to achieve low bypass for supersonic flight. When the valves are open, the bypass is increased to its maximum value for efficient subsonic flight.

Multiple flight-speed capability

Turboramjets. As noted above, the ramjet provides a simple and efficient means of propulsion for aircraft at relatively high supersonic flight speeds. It is, however, quite inefficient at transonic flight speeds and is completely ineffective at subsonic velocities. The turboramjet, shown in Figure 38, has been developed to overcome this inadequacy. In this system, a turbofan engine is built into the inlet of a ramjet engine to charge the latter with a pressurized stream of air at subsonic flight speed where ram pressure is insufficient for effective ramjet operation. During supersonic flight the fan blades, if they are of variable pitch, may be feathered so that they do not interfere with the flow of ram air to the ramjet. A separate inlet to the core engine that drives the fan may be closed off so as not to expose the turbomachinery to the hostile environment of the high-temperature ram air.

Built-in turbofan engine

Another variation of the turboramjet does without the core inlet and the core compressor altogether. Instead, the aircraft carries a tank of an oxidizer, such as liquid oxygen. The oxidizer is fed into the core combustion chamber along with the fuel to support the combustion process, which generates the hot gas stream to power the turbine that drives the fan. During supersonic flight, the fan may be feathered, and a surplus of fuel may be introduced into the core combustor. The unburned fuel passes through the fan turbine and undergoes combustion in the ramjet burner when it mixes with the fresh air entering via the bypass stream from the fan. (F.F.E.)

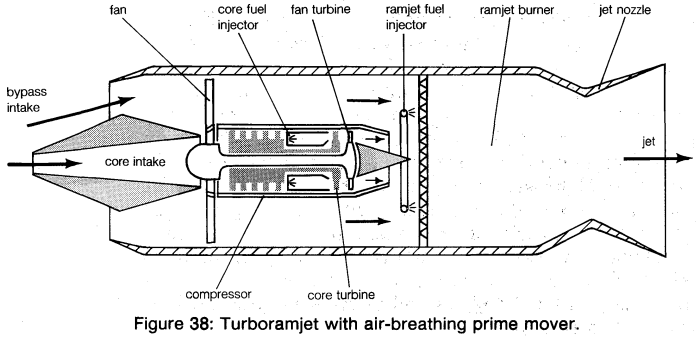


Figure 38: Turboramjet with air-breathing prime mover.

Development of jet engines. Like many other inventions, jet engines were envisaged long before they became a reality. The earliest proposals were based on adaptations of piston engines and were usually heavy and complicated. The first to incorporate a turbine design was conceived as early as 1921, and the essentials of the modern turbojet were contained in a patent in 1930 by Frank Whittle in England. His design was first tested in 1937 and achieved its first flight in May 1941. In Germany, parallel but completely independent work followed issuance of a patent in 1935. It proceeded more rapidly, and the very first flight of a turbojet-powered aircraft, a Heinkel HE-178, came in August 1939. By the end of World War II these prototype aircraft had developed into a few operational turbojet squadrons in the German, British, and U.S. air forces.

In the military area, jet fighter aircraft developed rapidly and were in use during the Korean War (1950–53), flying at speeds of 1,000 kilometres per hour. During the next decade they overcame the sound barrier and established normal operations up to more than twice the speed of sound (Mach 2). Bomber and transport jet aircraft were also able to reach and cruise at supersonic speeds.

Emergence of supersonic jet aircraft

The first civil jet transport, the British de Havilland Comet, flew in 1949, and regular transatlantic jet services were started in 1958 with the Comet 4 and the American Boeing 707. By 1974 more than 90 percent of hours flown throughout the world were flown by jets; the first supersonic airliner, the British-French Concorde, flying at more than twice the speed of sound, entered regular service in January 1976.

During the 1980s various major aircraft manufacturers undertook programs to develop fuel-saving propfan and unducted-fan propulsion systems. Some authorities believe that the next generation of commercial air transport may very well be powered by such advanced-technology propeller engines. (A.D.B./F.F.E.)

ROCKETS

General characteristics and principles of operation. The rocket constitutes a form of jet propulsion (see *Jet engines* above). It differs from the turbojet and other "air-breathing" engines in that all of the exhaust jet consists of the gaseous combustion products of "propellants" carried on board. Like the turbojet engine, the rocket develops thrust by the rearward ejection of mass at very high velocity.

The fundamental physical principle involved in rocket propulsion was formulated by Newton. According to his third law of motion, the rocket experiences an increase in momentum proportional to the momentum carried away in the exhaust,

$$M\Delta v_R = \mathring{m} v_e \Delta t = F\Delta t, \qquad (1)$$

where M is the rocket mass, Δv_R is the increase in velocity of the rocket in a short time interval, Δt, $\mathring{m}$ is the rate of mass discharge in the exhaust, v_e is the exhaust velocity (relative to the rocket), and F is force. The quantity $\mathring{m} v_e$ is the propulsive force, or thrust, produced on the rocket by exhausting the propellant,

$$F = \mathring{m} v_e . \qquad (2)$$

Evidently thrust can be made large by using a high mass discharge rate or high exhaust velocity. Employing high $\mathring{m}$ uses up the propellant supply quickly (or requires a large supply), and so it is preferable to seek high values of v_e. The value of v_e is limited by practical considerations, determined by how the exhaust is accelerated in the engine and what energy supply is available for the purpose.

Most rockets derive their energy in thermal form by combustion of condensed-phase propellants at elevated

Combustion of chemical propellants in conventional rockets

pressure. The gaseous combustion products are exhausted through a nozzle that converts part of the thermal energy to kinetic energy. The maximum amount of energy available is limited to that provided by combustion or by practical considerations imposed by the high temperature involved. Higher energies are possible if other energy sources (*e.g.*, electric arc or microwave heating) are used in conjunction with the chemical propellants on board the rockets, and extremely high energies are achievable when the exhaust is accelerated by electromagnetic means. As yet, these more exotic systems have not found application because of technical reasons but probably will be used in some future space missions where requisite electrical power sources can be shared by propulsion and other mission requirements (see *Other systems* below).

The exhaust velocity is a figure of merit for rocket propulsion because it is a measure of thrust per unit mass of propellant consumed—*i.e.*,

$$\frac{F}{\dot{m}} = v_e. \tag{3}$$

Values of v_e are in the range 2,000 to 5,000 metres per second for chemical propellants, while values two or three times that are claimed for electrically heated propellants. Values up to 40,000 metres per second are predicted for systems using electromagnetic acceleration. In engineering circles, notably in the United States, the exhaust velocity is widely expressed in units of pound thrust per pound weight per second, which is referred to as specific impulse. (In the International System of Units [SI], the unit of specific impulse is newton-seconds per kilogram.) Values in the range 185 to 465 seconds are analogous to the range of exhaust velocities noted above for chemical propellants.

Specific impulse

In a typical chemical-rocket mission, anywhere from 50 to 95 percent or more of the takeoff mass is propellant. This can be put in perspective by the equation for burnout velocity (gravity-free flight),

$$v_b = v_e \ln \frac{M_o}{M_s + M_{\text{pay}}}$$

$$= v_e \ln \frac{1}{\left(\frac{M_s}{M_p}\right)\left(\frac{M_p}{M_o}\right) + \left(\frac{M_{\text{pay}}}{M_o}\right)}. \tag{4}$$

In this expression, M_s/M_p is the ratio of propulsion system and structure weight to propellant weight, with a typical value of 0.09 (the symbol ln represents natural logarithm). M_p/M_o is the ratio of propellant weight to all-up takeoff weight, with a typical value of 0.90. A typical value for v_e for a hydrogen–oxygen system is 3,536 metres per second. From the above equation, the ratio of payload mass to takeoff mass (M_{pay}/M_o) can be calculated. For a low Earth orbit, v_b is about 7,544 metres per second, which would require M_{pay}/M_o to be 0.0374. In other words, it would take a 1,337,000-kilogram takeoff system to put 50,000 kilograms in a low orbit around the Earth. This is an optimistic calculation because equation (4) does not take into account the effect of gravity, drag, or directional corrections during ascent, which would double the takeoff mass. From equation (4) it is evident that there is a direct trade-off between M_s and M_{pay}, so that every effort is made to design for low structural mass, and M_s/M_p is a second figure of merit for the propulsion system. While the various mass ratios chosen depend strongly on the mission, rocket payloads generally represent a small part of the takeoff weight.

Multiple staging

A technique called multiple staging is used in many missions to minimize the size of the takeoff vehicle. A launch vehicle carries a second rocket as its payload, to be fired after burnout of the first stage (which is left behind). In this way, the inert components of the first stage are not carried to final velocity, with the second-stage thrust being more effectively applied to the payload. Most spaceflights use at least two stages. The strategy is extended to more stages in missions calling for very high velocities. The U.S. Apollo manned lunar missions used a total of six stages.

The unique features of rockets that make them useful include the following:

1. Rockets can operate in space as well as in the atmosphere of the Earth.
2. They can be built to deliver very high thrust (a modern heavy space booster has a takeoff thrust approaching 4.5 million kilograms).
3. The propulsion system can be relatively simple.
4. The propulsion system can be kept in a ready-to-fire state (important in military systems).
5. Small rockets can be fired from a variety of launch platforms, ranging from packing crates to shoulder launchers to aircraft (there is no recoil).

These features explain not only why all speed and distance records are set by rocket systems (air, land, space) but also why rockets are the exclusive choice for spaceflight. They also have led to a transformation of warfare, both strategic and tactical. Indeed, the emergence and advancement of modern rocket technology can be traced to weapon developments during and since World War II, with a modest but growing portion being funded through "space agency" initiatives such as the Ariane, Apollo, and Space Shuttle programs.

Chemical rockets. Rockets that employ chemical propellants come in different forms, but all share analogous basic components. These are (1) a combustion chamber where condensed-phase propellants are converted to hot gaseous reaction products, (2) a nozzle to accelerate the gas to high exhaust velocity, (3) propellant containers, (4) a means of feeding the propellants into the combustion chamber, (5) a structure to support and protect the parts, and (6) various guidance and control devices.

Basic parts of a chemical rocket

Chemical rocket propulsion systems are classified into two general types according to whether they burn solid or liquid propellants. Solid systems are usually called motors and liquid systems are referred to as engines. Some developmental work has been carried out on so-called hybrid systems, in which the fuel is a solid and the oxidizer is a liquid, or vice versa. The characteristics of such systems differ greatly depending on the requirements of a given mission.

Solid-rocket motors. The principal features of a solid-rocket motor (SRM) are shown in Figure 39. The propellant consists of one or more pieces mounted directly in the motor "case," which serves both as a propellant tank and combustion chamber. The propellant is usually arranged to protect the motor case from heating. Most modern propellant charges are formed by pouring a viscous mix into the motor case with suitable mold fixtures. The propellant solidifies (usually by polymerization) and the mold fixtures are removed, leaving the propellant bonded to the motor case with a suitably shaped perforation down the middle. During operation the solid burns on the exposed surfaces. These burn away at a predictable rate to give the desired thrust.

The motor case generally consists of a steel or aluminum tube; it has a head-end dome that contains an igniter

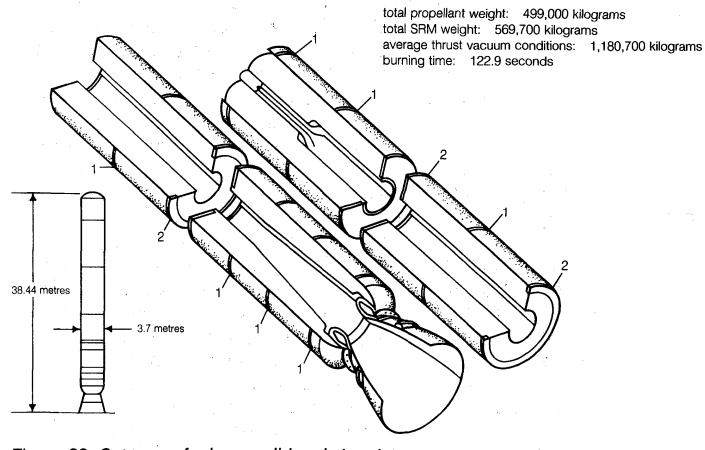


Figure 39: *Cutaway of a large solid-rocket motor.*
This type of motor, used on the U.S. Space Shuttle, consists of four segments and a nozzle assembly that are mated together at the launch site. The numbers indicate joints in the steel case: (1) "factory joints," which are case-segment joints assembled before propellant casting, and (2) "field joints," which are assembled subsequently. The Shuttle motors are recovered at sea, refurbished, and reused.

and an aft-end dome that houses or supports the nozzle. Motor cases ordinarily have insulation on their interior surfaces, especially those not covered by propellant, for protection against thermal degradation. When a mission requires particularly lightweight components, motor cases are often made by filament winding of high-strength fibres on a suitable form. The filaments are held in place by continuous application and curing of plastic during winding. In motor cases, the front and aft domes are wound as integral parts of the case, with suitable openings and fixtures included to permit removal of the (collapsible) motor case form, loading of propellant, and attachment of igniter and nozzle. No matter what type of motor construction is involved, provisions must be made for attaching the structures that connect to the rest of the vehicle and to the launching pad or vehicle. In nearly all applications, the motor case constitutes the main structural component of the rocket and must be designed accordingly.

Propellants for solid-rocket motors are made from a wide variety of substances, selected for low cost, acceptable safety, and high performance. The selection is strongly affected by the specific application. Typical ingredients are

Propellant materials

ammonium perchlorate (a granular oxidizer), powdered aluminum (a fuel), and polybutadiene-acrylonitrile-acrylic acid (a fuel that is liquid during mixing and that polymerizes to a rubbery binder during curing). This combination is used in major U.S. space boosters (*e.g.*, the Space Shuttle and the Titan). Higher performance is achieved by the use of more energetic oxidizers (*e.g.*, cyclotetramethylene tetranitramine [HMX]) and by energetic plasticizers in the binder or by energetic binders such as a nitrocellulose–nitroglycerin system. In military systems, low visibility of the exhaust plume has sometimes been a requirement, which precludes the use of aluminum powder or very much ammonium perchlorate and makes it necessary to use other materials such as HMX and high-energy binder systems that yield combustion products involving mainly carbon, oxygen, hydrogen, and nitrogen.

Propellant charges must meet a variety of often conflicting requirements. From a performance standpoint, they should burn inward at the burning surface in a consistent and predictable manner that is not unduly sensitive to pressure or bulk temperature at a rate typically in the range of 0.2 to 20 centimetres per second. They should be as dense as possible (to maximize the amount of propellant in a given motor size) while still producing reaction products of low molecular weight and high temperature (to maximize exhaust velocity). From a practical standpoint, propellants must be insensitive to accidental ignition stimuli and amenable to safe manufacturing and loading in the motor. Once they have been loaded in the motor, they must achieve and retain the mechanical properties necessary to maintain structural integrity under shipping, storage, and flight conditions. Since the energetic materials used in high-performance propellants are often explosives, manufacturing the propellant to a safe form is a complex technology involving special facilities and strict safety guidelines. To a degree this is true also of less sensitive propellants (*e.g.*, ammonium perchlorate–aluminum–polymeric binder propellants) used in intermediate-performance systems, such as the Space Shuttle booster motors.

Nozzle

The principal requirement for a nozzle is that it be able to produce an optimum flow of the exhaust gas from combustion chamber pressure to exterior pressure (or thereabouts), a function that is accomplished by proper contouring and sizing of the conduit. The contour is initially convergent to a "throat" section. The velocity of the exhaust gas in this region is equal to the local velocity of sound, and the throat cross-sectional area controls the mass discharge rate (and hence the operating pressure). Beyond the throat, the channel is divergent and the flow accelerates to high supersonic speeds with a corresponding pressure decrease. Contours are often carefully designed so that shock waves do not form. (Shock waves slow the flow and degrade thrust.)

The details of nozzle design depend strongly on application. Most applications require at least some use of insulation or special high-temperature materials (*e.g.*, graphite) in order to protect the load-carrying structures from thermal degradation. Many applications require that the direction of the exhaust flow be controllable over a few degrees in order to provide for "steering." This is accomplished in a variety of ways that frequently complicate the design considerably and increase nozzle weight.

The ignition system

The igniter in a solid-rocket motor provides a means of heating the surface of the propellant charge to a high enough temperature to induce combustion. At the same time, the igniter is usually designed to produce some initial pressure increase in the motor to assure more reproducible start-up. The igniter consists of a container of material like a metal–oxidizer mixture that is more easily and quickly ignited than the propellant; it is initiated by an electric squib or other externally energized means. The igniter case is designed to be sealed until fired and to disperse hot and burning products when pressurized by its own burning. In large motors the igniter may feed into a miniature motor containing a fast-burning propellant charge, which exhausts into the main motor to produce ignition and pressurization. Most ignition systems include some kind of "arming" feature that prevents ignition by unintended stimuli.

Thrust levels of solid rockets

The thrust level of a solid rocket is determined by the rate of burning of the propellant charge (mass rate in equation [2]), which is determined by the surface area (S_c) that is burning and the rate (r) at which the surface burns into the solid. The designer chooses a charge geometry that will vary with time during burning in the manner needed for a particular mission and chooses a propellant formulation that gives the desired burning rate. This means that the thrust-time function is not amenable to much intentional modification after manufacture, and most missions using solid-rocket motors are designed to take advantage of the predictability of the thrust-time function rather than to regulate thrust during flight. The lack of real-time control on thrust is compensated for by the ability to achieve extraordinarily high mass-flow rates without the propellant pumps ordinarily used in liquid-propellant rockets. The thrust levels occurring in practice depend on motor operating pressure, which in turn is shown in internal ballistic theory to depend on motor and propellant properties according to the equation

$$p = \left(\rho_p \frac{C}{C_d} \frac{S_c}{A_t}\right)^{1/(1-n)}, \tag{5}$$

where A_t is nozzle throat area, C_d is a nozzle discharge coefficient (that depends on the thermochemical properties of the propellant reaction products), ρ_p is the density of the solid propellant, and C and n are constants in an equation that gives the approximate dependence of burning rate of the propellant on pressure,

$$r = Cp^n. \tag{6}$$

The thrust is then given by an engineering equation,

$$F = C_F A_t p = C_F \left(\rho_p \frac{CS_c}{C_d A_t}\right)^{1/(1-n)} A_t, \tag{7}$$

where C_F depends on nozzle geometry, thermochemical properties, and to a lesser degree on external pressure. Typical values of the quantities in this equation are given in Table 1.

Table 1: Typical Values of Internal Ballistic Variables in Equation (7)

ρ_p	C_d	S_c/A_t	C	n	C_F	coefficient of $A_t(C_F p)$
kg/m^3	s/m	—	$N^{-1}m^{1+2n}s^{-1}$	—	—	N/m^2
1,762	6.1×10^{-4}	200	4.45×10^{-5}	0.35	1.5	8.21×16^6

In most applications, the need to minimize the mass of motor components is a major design consideration. This need is so important that it is often "bought" at the expense of low safety margins and sometimes by the use of exotic construction and structural materials. These considerations are constantly weighed against the cost of mission failures. With the advent of manned flight and payloads sometimes costing $1 billion or more, the thinking on safety margins and acceptable propulsion-system cost is changing.

Liquid-propellant rocket engines. Liquid-propellant systems carry the propellant in tanks external to the combustion chamber. Most of these engines use a liquid oxidizer and a liquid fuel, which are transferred from their respective tanks by pumps. The pumps raise the pressure well above the operating pressure of the engine, and the propellants are then injected into the engine in a manner that assures atomization and rapid mixing. Liquid-propellant engines have certain features that make them preferable to solid systems in many applications. These features include (1) higher attainable exhaust velocities (v_e), (2) higher mass fractions (propellant mass divided by mass of inert components), and (3) control of operating level in flight (throttleability), sometimes including stop-and-restart capability and emergency shutdown. Also, in some applications it is an advantage that propellant loading is delayed until shortly before launch time, a measure that the use of a liquid propellant allows. These features tend to promote the use of liquid systems in many upper-stage applications where high v_e and high propellant mass fraction are particularly important. Liquid systems also have been used extensively as first-stage launch vehicles for space missions, as, for example, in the Saturn (U.S.), Ariane (European), and Energia (Soviet) launch systems. Many intercontinental ballistic missile (ICBM) systems employ liquid-propellant engines, but solid systems have been widely adopted for these applications in the United States because of their suitability for launch on short notice. The relative merits of solid and liquid propellants in heavy launch vehicles are still under debate and involve not only propulsion performance but also issues related to logistics, capital and operating costs of launch sites, recovery and reuse of flight hardware, and so forth.

Distinctive features

The typical components of a liquid-rocket propulsion system are the engine, fuel tanks, and vehicle structure with which to hold these parts in place and connect to payload and launch pad (or vehicle). The fuel and oxidizer tanks are usually of very lightweight construction, as they operate at low pressure. In some applications, the propellants are cryogenic (*i.e.,* they are substances like oxygen and hydrogen that are gaseous at ambient temperature and must be tanked at very low temperature to be in the liquid state).

Principal components

The liquid-propellant engine itself (Figure 40) consists of a main chamber for mixing and burning the fuel and oxidizer, with the fore end occupied by fuel and oxidizer manifolds and injectors and the aft end comprised of the nozzle. Integral to the main chamber is a coolant jacket through which liquid propellant (usually fuel) is circulated at rates high enough to allow the engine to operate continuously without an excessive increase of temperature in the chamber. Engine operating pressures are usually in the range 1,000 to 10,000 kilopascals (10 to 100 atmospheres). The propellants are supplied to the injector manifold at a somewhat higher pressure, usually by high-capacity turbopumps (one for the fuel and another for the oxidizer). From the outside, a liquid-propellant engine often looks like a maze of plumbing, which connects the tanks to the pumps, carries the coolant flow to and from the cooling jackets, and conveys the pumped fluids to the injector. In addition, engines are generally mounted on gimbals so that they can be rotated a few degrees for thrust direction control, and appropriate actuators are connected between the engine (or engines) and the vehicle structure to constrain and rotate the engine.

Each of the main engines of the U.S. Space Shuttle (shown in Figure 40) employs liquid oxygen (LO_2) and liquid hydrogen (LH_2) propellants. These engines represent a very complex, high-performance variety of liquid-propellant rocket. Not only does each have a v_e value of 3,630 metres per second but is also capable of thrust-magnitude control over a significant range (2–1). Moreover, the Shuttle engines are part of the winged orbiter, which is designed to carry both crew and payload for up to 20 missions.

Main engines of the Space Shuttle

At the opposite extreme of complexity and performance is a hydrazine thrustor used for attitude control of conventional flight vehicles and unmanned spacecraft. Such a system may employ a valved pressure vessel in place of a pump, and the single propellant flows through a catalyst bed that causes exothermic (heat-releasing) decomposition. The resulting gas is exhausted through a nozzle that is suitably oriented for the required attitude correction. Systems of this kind also are used as gas generators for turbopumps on larger rockets.

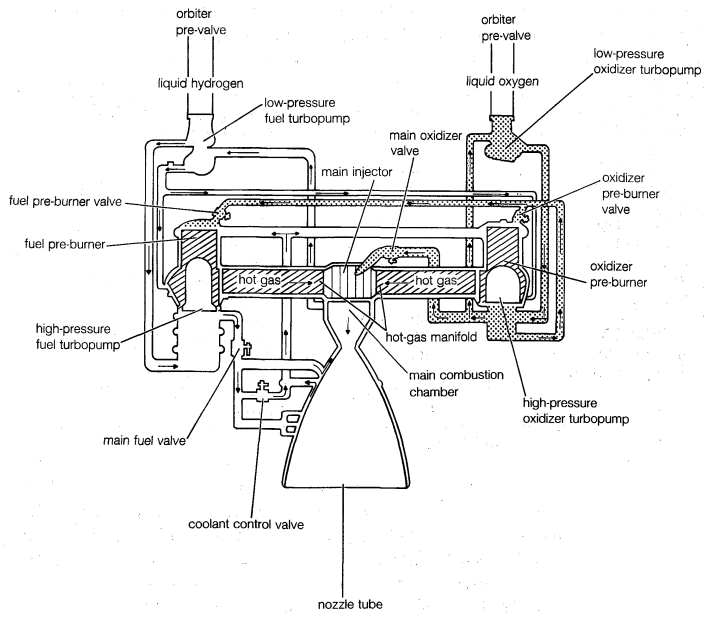


Figure 40: Flow diagram for the Space Shuttle main engine (SSME). Three such engines are mounted on the orbiter.

Most liquid-propellant rockets use bipropellant systems—*i.e.,* those in which an oxidizer and a fuel are tanked separately and mixed in the combustion chamber. Desirable properties for propellant combinations are low molecular weight and high temperature of reaction products (for high exhaust velocity), high density (to minimize tank weight), low hazard factor (*e.g.,* corrosivity and toxicity), and low cost. Choices are based on trade-offs according to the applications. For example, liquid oxygen is widely used because it is a good oxidizer for a number of fuels (giving high flame temperature and low molecular weight) and because it is reasonably dense and relatively inexpensive. It is liquid only below −183° C, which somewhat limits its availability, but it can be loaded into insulated tanks shortly before launch (and replenished or drained in the event of launch delays). Liquid fluorine or ozone are better oxidizers in some respects but involve more hazard and higher cost. The low temperatures of all of these systems require special design of pumps and other components, and the corrosivity, toxicity, and hazardous characteristics of fluorine and ozone have apparently thus far prevented their use in operational systems. Other oxidizers that have seen operational use are nitric acid (HNO_3) and nitrogen tetroxide (N_2O_4), which are liquids under ambient conditions. While they are somewhat noxious chemicals, they are useful in applications where the rocket must be in a near ready-to-fire condition over an extended period of time, as in the case of long-range ballistic missiles.

Predominance of bipropellant systems

Liquid hydrogen is usually the best fuel from the standpoint of high exhaust velocity, and it might be used exclusively were it not for the cryogenic requirement and its low density. Such hydrocarbon fuels as alcohol and kerosene are often preferred because they are liquid under ambient conditions and denser than liquid hydrogen in addition to being more "concentrated" fuels (*i.e.,* they have more fuel atoms in each molecule). The values of exhaust velocity are determined by the relative effects of higher flame (combustion) temperatures and molecular weights of reaction products (as compared to liquid oxygen and liquid hydrogen).

Hydrocarbon fuels

In practice, a variety of choices of propellant systems have been made in major systems, as shown in Table 2. In flights where cryogenic propellants can be utilized (*e.g.,* ground-to-earth-orbit propulsion), liquid oxygen is usually

Table 2: Liquid Propellants in Various Flight Vehicles

rocket	oxidizer	fuel
German V-2	liquid oxygen	ethyl alcohol–water (75%–25%)
Atlas ICBM	liquid oxygen	RP-1 (kerosene)
Delta		
First stage	liquid oxygen	RP-1 (kerosene)
Second stage	nitrogen tetroxide	hydrazine–UDMH* (50%–50%)
Saturn		
First stage	liquid oxygen	RP-1 (kerosene)
Second stage	liquid oxygen	liquid hydrogen
Third stage	liquid oxygen	liquid hydrogen
Apollo Lunar Module	nitrogen tetroxide	hydrazine–UDMH* (50%–50%)
Space Shuttle		
Main engines	liquid oxygen	liquid hydrogen
Orbital Maneuvering System	nitrogen tetroxide	monomethyl hydrazine
Ariane 4, first stage	nitrogen tetroxide	UDMH*
Energia, first stage		
Core	liquid oxygen	liquid hydrogen
Cluster	liquid oxygen	kerosene

*Unsymmetrical dimethylhydrazine.

used as the oxidizer. In first stages either a hydrocarbon or liquid hydrogen is employed, while the latter is usually adopted for second stages. In ICBMs and other similar guided missiles that must stand ready for launch on short notice, noncryogenic (or "storable") propellant systems are used, as, for instance, an oxidizer–fuel mixture of nitrogen tetroxide and hydrazine–unsymmetrical dimethylhydrazine (also designated UDMH; $[CH_3]_2\ NNH_2$). Systems of this sort also find application on longer duration flights such as those involving the Space Shuttle Orbital Maneuvering System and the Apollo Lunar Module. Solid motors have proved useful on long-duration flights, but liquid systems are often preferred because of the need for stop–start capability or thrust control.

Possible use of energy sources independent of the propellant fluid

Other systems. As suggested earlier, systems using energy sources independent of the propellant fluid have been studied, and they offer promise for some future missions. In certain systems the propellant is heated at elevated pressure by independent means and then accelerated by exhaust through a nozzle. In others the propellant is accelerated by electromagnetic means, in which case at least part of the fluid must be electrically charged first. In these systems the energy source may be nuclear, solar, or beamed energy from an independent source. The outlook for most current missions is that on-board energy sources of this kind would be too heavy, especially for high-thrust missions. There are, however, missions such as manned flights to other planets where sustained low thrust from on-board energy sources would shorten mission duration greatly, saving both time and consumable materials. Such a mission would very likely originate from Earth orbit, with flight system and on-board materials being transported to Earth orbit by chemical rocket propulsion. Electrically heated fluids would probably be used in missions involving manned space stations, where low-thrust capability is needed to control orbit and station attitude. Consideration is even being given to the use of waste products as propellants; these could be heated electrically from power systems already on board for station operational needs.

Use of black powder as a propellant

Development of rockets. The technology of rocket propulsion appears to have its origins in the period AD 1200–1300 in Asia, where the first "propellant" (a mixture of saltpetre, sulfur, and charcoal called black powder) had been in use for about 1,000 years for other purposes. As is so often the case with the development of technology, the early uses were primarily military. Powered by black powder charges, rockets served as bombardment weapons, culminating in effectiveness with the Congreve rockets (named for William Congreve, a British officer who was instrumental in their development) of the early 1800s. Performance of these early rockets was poor by modern standards because the only available propellant was black powder, which is not ideal for propulsion. Military use of rockets declined from 1815 to 1936 because of the superior performance of guns.

During the period 1880–1930 the idea of using rockets for space travel grew in public interest. Stimulated by the conceptions of such fiction writers as Jules Verne, the Russian scientist Konstantin E. Tsiolkovsky worked on theoretical problems of propulsion-system design and rocket motion and on the concept of multistage rockets. Perhaps more widely recognized are the contributions of Robert H. Goddard, an American scientist and inventor who from 1908 to 1945 conducted a wide array of rocket experiments. He independently developed ideas similar to those of Tsiolkovsky about spaceflight and propulsion and implemented them, building liquid- and solid-propellant rockets. His developmental work included tests of the world's first liquid-propellant rocket in 1926. Goddard's many contributions to the theory and design of rockets earned him the title of father of modern rocketry. A third pioneer, Hermann Oberth of Germany, developed much of the modern theory for rocket and spaceflight independent of Tsiolkovsky and Goddard. He not only provided inspiration for visionaries of spaceflight but played a pivotal role in advancing the practical application of rocket propulsion that led to the development of rockets in Germany during the 1930s.

The contributions of Tsiolkovsky, Goddard, and Oberth

Due to the work of these early pioneers and a host of rocket experimenters, the potential of rocket propulsion was at least vaguely perceived prior to World War II, but there were many technical barriers to overcome. Development was accelerated during the late 1930s and particularly during the war years. The most notable achievements in rocket propulsion of this era were the German liquid-propellant V-2 rocket and the Me-163 rocket-powered airplane. (Similar developments were under way in other countries but did not see service during the war.) A myriad of solid-propellant rocket weapons also were produced, and tens of millions were fired during combat operations by German, British, and U.S. forces (see WAR, THE TECHNOLOGY OF: *Rockets and missile systems*). The main advances in propulsion that were involved in the wartime technology were the development of pumps, injectors, and cooling systems for liquid-propellant engines and high-energy solid propellants that could be formed into large pieces with reliable burning characteristics.

From 1945 to 1955 propulsion development was still largely determined by military applications. Liquid-propellant engines were refined for use in supersonic research aircraft, intercontinental ballistic missiles (ICBMs), and high-altitude research rockets. Similarly, developments in solid-propellant motors were in the areas of military tactical rocket applications and high-altitude research. Bombardment rockets, aircraft interceptors, antitank weapons, and air-launched rockets for air and surface targets were among the primary tactical applications. Technological advances in propulsion included the perfection of methods for casting solid-propellant charges, development of more energetic solid propellants, introduction of new structural and insulation materials in both liquid and solid systems, manufacturing methods for larger motors and engines, and improvements in peripheral hardware (*e.g.*, pumps, valves, engine-cooling systems, and direction controls). By 1955 most missions called for some form of guidance, and larger rockets generally employed two stages. While the potential for spaceflight was present and contemplated at the time, financial resources were directed primarily toward military applications.

The next decade witnessed the development of large solid-propellant rocket motors for use in ICBMs, a choice motivated by the perceived need to have such systems in ready-to-launch condition for long periods of time. This resulted in a major effort to improve manufacturing capabilities for large motors, lightweight cases, energetic propellants, insulation materials that could survive long operational times, and thrust-direction control. Enhancement of these capabilities led to a growing role for solid-rocket motors in spaceflight. Between 1955 and 1965 the vision of the early pioneers began to be realized with the achievement of Earth-orbiting satellites and manned spaceflight. The early missions were accomplished with liquid-propulsion systems adapted from military rockets. The first successful "all-civilian" system was the Saturn launch vehicle for the Apollo Moon-landing program, which used five 680,000-kilogram-thrust liquid-propellant engines in the first stage. Since then, liquid systems have been employed by most countries for spaceflight applications, though solid boost-

ers have been combined with liquid engines in various first stages of U.S. launch vehicles (those of the Titan 34D, Delta, and Space Shuttle) and solid-rocket motors have been used for several systems for transfer from low Earth orbit to geosynchronous orbit. In such systems, the lower performance of solid-propellant motors is accepted in exchange for the operational simplicity that it provides.

Since 1965, missions have drawn on an ever-expanding technology base, using improved propellants, structural materials, and designs. Present-day missions may involve a combination of several kinds of engines and motors, each chosen according to its function. Because of the performance advantages of energetic propellants and low structural mass, propulsion systems are operated near their safe limits, and one major challenge is to achieve reliability commensurate with the value of the (sometimes human) payload. (E.W.P.)

Nuclear fission reactors

A nuclear reactor is a device in which a nuclear fission chain reaction takes place under controlled conditions. Such devices are used as research tools, as systems for producing radioisotopes, and most prominently as energy sources. The latter are commonly called power reactors.

Fission is the process in which a heavy nucleus splits into two smaller fragments. A large amount of energy is released in this process, and this energy is the basis of fission power systems. The nuclear fragments are in very excited states and emit neutrons and other forms of radiation. The neutrons can then cause new fissions, which in turn yield more neutrons, and so forth. Such a continuous self-sustaining series of fissions constitutes a fission chain reaction. For a detailed discussion of nuclear fission, see ATOMS: *Fundamentals of the fission process.*

In an atomic bomb the chain reaction is designed to increase in intensity until much of the material has fissioned. This increase is very rapid and produces the extremely sharp, tremendously energetic explosions characteristic of such bombs. In a nuclear reactor the chain reaction is maintained at a controlled, nearly constant level. Nuclear reactors are so designed that they cannot explode like atomic bombs.

Energy release

Most of the energy of fission—about 85 percent of it—is released within a very short time after the process occurs. The rest of the energy comes from the radioactive decay of fission products, which is what the fragments are called after they have emitted neutrons. Radioactive decay continues when the fission chain has been stopped, and its energy must be dealt with in any proper reactor design.

PRINCIPLES OF OPERATION

Chain reaction and criticality. The course of a chain reaction is determined by the probability that a neutron released in fission will cause a subsequent fission. If on the average less than one neutron causes another fission, the rate of fission will decrease with time and ultimately drop to zero. This situation is called subcritical. When an average of one neutron from a fission causes another fission, the fission rate is steady and the reactor is critical. A critical reactor is what is usually desired. When more than one neutron causes a subsequent fission, fission rate and power increase and the situation is termed supercritical. In order to be able to increase power, reactors are designed to be slightly supercritical when all controls are removed.

Reactivity

Reactor control. A parameter called reactivity is positive when a reactor is supercritical, zero at criticality, and negative when the reactor is subcritical. Reactivity can be controlled in various ways: by adding or removing fuel; by changing the fraction of neutrons that leaks from the system; or by changing the amount of an absorber that competes with the fuel for neutrons. Control is generally accomplished by varying absorbers, which are commonly in the form of movable elements—control rods—or sometimes by changing the concentration of the absorber in a reactor coolant. Leakage changes are usually automatic; for example, an increase of power may cause coolant to boil (see below), which in turn increases neutron leakage and reduces reactivity. This, and other types of negative power-reactivity feedbacks, are vital aspects of safe reactor design.

Reactor control is facilitated by the presence of delayed neutrons. These neutrons are emitted by fission products some time after fission has occurred. The fraction of delayed neutrons is small, but there is a sufficient number of such neutrons for the types of changes needed to regulate an operating reactor, and so the chain reaction must "wait" for them before it can respond. This eases operation considerably.

Fissile and fertile materials. All heavy nuclides can fission if they are in an excited enough state, but only a few fission readily when struck by slow (low-energy) neutrons. Such species of atoms are called fissile. The most important of these are uranium-233 (^{233}U), uranium-235 (^{235}U), plutonium-239 (^{239}Pu), and plutonium-241 (^{241}Pu). The only one that occurs in usable amounts in nature is uranium-235, which makes up a mere 0.711 percent of natural uranium by weight. Uranium-233 can be produced by neutron capture in natural thorium (^{232}Th); that is to say, when a nucleus of thorium-232 absorbs a neutron, it becomes uranium-233. Similarly, plutonium-239 is created by neutron capture in uranium-238 (^{238}U; the principal constituent of naturally occurring uranium), and plutonium-241 is formed when a neutron is absorbed into plutonium-240 (^{240}Pu). Plutonium-240 builds up over time in most power reactors. Thorium-232, uranium-238, and plutonium-240 are termed fertile materials because they can be transformed into fissile materials.

A power reactor contains both fissile and fertile materials. The fertile materials replace fissile materials that are destroyed by fission. This permits the reactor to run longer before the amount of fissile material decreases to the point where criticality can no longer be maintained.

Heat removal. The energy of fission is quickly converted to heat, the bulk of which is deposited in the fuel. A coolant is therefore required to remove this heat. The most common coolant is water, but any fluid can be used. Heavy water (deuterium oxide), air, carbon dioxide, helium, liquid sodium, sodium-potassium alloy (called NaK), molten salts, and hydrocarbons have all been used in reactors or reactor experiments. Some research reactors are operated at very low power and have no need for a dedicated cooling system; in such units the small amount of heat that is generated is removed by conduction and convection to the environment. Very high power reactors must have extremely sophisticated cooling systems to remove heat quickly and reliably; otherwise, the heat will build up in the reactor fuel and melt it.

Coolant

Shielding. An operating reactor is a powerful source of radiation, since fission and subsequent radioactive decay produce neutrons and gamma rays, both of which are highly penetrating radiations. A reactor must have special shielding around it to absorb this radiation in order to protect technicians and other reactor personnel. In a popular class of research reactors known as "swimming pools," this shielding is provided by placing the reactor in a large, deep pool of water. In other kinds of reactors, the shield consists of a thick concrete structure around the reactor system. The shield also may contain heavy metals, such as lead or steel, for more effective absorption of gamma rays, and heavy aggregates may be used in the concrete itself for the same purpose.

Critical concentration and size. Not every arrangement of material containing fissile fuel can be brought to criticality. Even if there were no leakage of neutrons from a reactor, a critical concentration of fissile material must be present. Otherwise, absorption of neutrons by other constituents of the reactor will be too high to permit a critical chain reaction to proceed. Similarly, even if there is a high enough concentration for criticality, the reactor must be large enough so that not too many neutrons leak out before being absorbed. This imposes a critical size limit on a reactor of a given concentration.

Although the only useful fissile material in nature, uranium-235, is found in natural uranium, there are just a few combinations and arrangements of this and other materials that can be brought to criticality. To increase the range of feasible reactor designs, enriched uranium can

Use of enriched uranium fuel

be used. Most of today's power reactors employ enriched uranium fuel in which the percentage of uranium-235 has been increased to 3 to 4 percent. This is about five times the concentration in natural uranium. Large plants for enriching uranium exist in several countries; enrichment has now become a commercial enterprise (see below).

Thermal, intermediate, and fast reactors. Reactors are conveniently classified according to the typical energies of the neutrons that cause fission. Neutrons emanating in fission are very energetic; their average energy is around two million electron volts (MeV), 80 million times higher than the energy of atoms in ordinary matter at room temperature. As the neutrons collide with nuclei in a reactor, they lose energy. The choice of reactor materials and of fissile material concentrations determines how much they are slowed down by these collisions before causing fission.

In a thermal reactor, enough collisions are permitted to occur so that most of the neutrons reach thermal equilibrium with the atoms of the reactor at energies of a few hundredths of an electron volt. Neutrons lose energy most efficiently by colliding with light atoms such as hydrogen (mass 1), deuterium (mass 2), beryllium (mass 9), and carbon (mass 12). Materials that contain atoms of this kind—water, heavy water, beryllium metal and oxide, and graphite—are deliberately incorporated into the reactor for this reason and are known as moderators. Since water and heavy water also can function as coolants, they can do double duty in thermal reactors.

Moderators

One disadvantage of thermal reactors is that at low energies uranium-235 and plutonium-239 not only can be fissioned by thermal (or slow) neutrons but also can capture neutrons without undergoing fission. This destroys fissile atoms without any fission to show for it. When neutrons of higher energy cause fission, fewer of these captures occur. To achieve this, a reactor can be built to operate without a moderator. Then, depending on how many collisions take place with heavier atoms before fission occurs, the typical fission-causing neutrons can have energies in the range of 0.5 electron volt to thousands of electron volts (intermediate reactors) or several hundred thousand electron volts (fast reactors). Such reactors require higher concentrations of fissile material to reach criticality than do thermal reactors but are more efficient at converting fertile material to fissile material. Indeed, they can be designed to produce more than one new fissile atom for each fissile atom destroyed. Such reactors are called breeders. Breeder reactors may become particularly important if the world demand for nuclear power turns out to be a long-term one, since their fuel is manufactured from very abundant fertile materials.

Breeder reactors

REACTOR DESIGN AND COMPONENTS

There are a large number of ways in which a reactor may be designed and constructed, and many types have been experimentally realized. Over the years, nuclear engineers have developed reactors with solid fuels and liquid fuels, thick reflectors and no reflectors, forced cooling circuits and natural conduction or convection heat-removal systems, and so on. Most reactors, however, have certain basic components. These are described below.

Core. All reactors have a core, a central region that contains the fuel, fuel cladding, coolant, and, where separate from the latter, moderator. It is in the core that fission occurs and the resulting neutrons migrate.

Fuel elements

The fuel is usually heterogeneous—*i.e.*, it consists of elements containing fissile material along with a diluent. This diluting agent may be fertile material or simply material that has good mechanical and chemical properties and that does not readily absorb neutrons. The diluted fissile material is enclosed in a cladding—a substance that isolates the fuel from the coolant and keeps the radioactive fission products contained.

Fuel types. Different kinds of reactors use different types of fuel elements. For example, the light-water reactor (LWR), which is the most widely used variety for commercial power generation in the United States, employs a fuel consisting of pellets of sintered uranium dioxide loaded into cladding tubes of zirconium alloy that measure about one centimetre in diameter and roughly three to four metres long. These tubes, called pins, are bundled together into a fuel assembly, with the pins arranged in a square lattice. The uranium used in the fuel is 3- to 4-percent enriched. Since light (ordinary) water tends to absorb more neutrons than do other moderators, such enrichment is crucial. The CANDU (Canadian deuterium-uranium) reactor, which is the principal type of heavy-water reactor, uses natural uranium compacted into pellets. These pellets are inserted in tubes arranged in a lattice. Such a fuel assembly measures about one metre in length, and several assemblies are arranged end-to-end within a channel inside the reactor core.

In a high-temperature graphite reactor the fuel is made of small spherical particles containing uranium dioxide at the centre with concentric shells of carbon, silicon carbide, and carbon around them. (These shells serve as microscopic cladding.) The particles are mixed with graphite and encased in a macroscopic graphite cladding. In a sodium-cooled fast reactor, commonly called a liquid-metal reactor (LMR), the fuel consists of dioxide pellets (French design) or uranium-plutonium-zirconium metal alloy pins (U.S. design) in steel cladding.

The most common type of fuel used in research reactors consists of plates of a uranium-aluminum alloy with an aluminum cladding. The uranium is enriched to 20 percent, and silicon, along with aluminum, are included in the "meat" of the plate. A common variety of research reactor, known as TRIGA (from training, research, and isotope-production reactors–General Atomic), employs a fuel of mixed uranium and zirconium hydride in zirconium cladding.

Coolants and moderators. A variety of substances, including light water, heavy water, air, carbon dioxide, helium, liquid sodium, liquid sodium-potassium alloy, and hydrocarbons (oils), have been used as coolants. Such substances are good conductors of heat and serve to carry the thermal energy produced by fission from the core to the steam-generating equipment of the nuclear power plant.

In many cases, the same substance functions as both coolant and moderator, as in the case of light and heavy water. The moderator slows down the fast (high-energy) neutrons emitted in fission to speeds at which they are more likely to induce fission. In doing so, the moderator helps initiate and sustain a fission chain reaction.

Reflector. A reflector is a region of unfueled material surrounding the core. Its function is to scatter neutrons that leak from the core and thereby return some of them to the core. This reduces core size and smooths out the power density. The reflector is particularly important in research reactors, since it is the region in which much of the experimental apparatus is located. Some reflectors are located inside the core as central islands in which high neutron intensities can be achieved for experimental purposes. In most types of power reactors, a reflector is less important, because the reactors are large and do not leak many neutrons. Yet, as it serves to keep the power density uniform, such an unfueled zone of moderator material is left around the core. The liquid-metal reactor represents a special case. Most sodium-cooled reactors are deliberately built to allow a large fraction of their neutrons—those not needed to maintain the chain reaction—to leak from the core. These neutrons are valuable because they can produce new fissile material if they are absorbed by fertile material. Thus, fertile material—generally depleted uranium or its dioxide—is placed around the core to catch the leaking neutrons. Such an absorbing reflector is referred to as a blanket or a breeding blanket.

Reducing the leakage of neutrons

Reactor control elements. All reactors need special elements for control. Although control can be achieved by varying parameters of the coolant circuit or by varying the amount of absorber dissolved in the coolant or moderator, by far the most common method involves the use of special absorbing assemblies—namely, control rods or sometimes blades. Typically a reactor is equipped with three types of rods for different purposes: (1) safety rods for starting up and shutting down the reactor, (2) regulating rods for adjusting the reactor's power rate, and (3) shim rods for compensating for changes in reactivity as fuel is depleted by fission and capture.

Control rods

The most important function of the safety rods is to shut down the reactor, either when such a shutdown is scheduled or in case of a real or suspected emergency. These rods contain enough absorber to terminate a chain reaction under any conceivable condition. They are withdrawn before fuel is loaded and remain available in case a loading error requires their action. After the fuel is loaded, the rods are inserted, to be withdrawn again when the reactor is ready for operation. The mechanism by which they are moved is designed to be fail-safe in the sense that if there is a mechanical failure the safety rods will fall by gravity into the reactor. In some cases, moreover, the safety rods have an automatic feature, such as a fuse, which releases them by virtue of physical effects independent of electronic signals.

Regulating rods are deliberately designed to affect reactivity only by a small degree. It is assumed that at some time the rods might be totally withdrawn by mistake, and the idea is to keep the added reactivity in such cases well within sensible limits. A well-designed regulating rod will add so little reactivity when it is removed that the delayed neutrons will continue to control the rate of power increase.

Shim rods are designed to compensate for the effects of burnup (*i.e.,* energy production). Reactivity changes resulting from burnup can be large, but they occur slowly over periods of days to years, as compared to the seconds-to-minutes range over which safety actions and routine regulation take place. Therefore, shim rods may control a significant amount of reactivity, but they will work perfectly well under constraints on their speed of movement. A common way in which shims are operated is by inserting or removing them as regulating rods reach the end of their most useful position range. When this happens, shim rods are moved so that the regulating rods can be reset.

The functions of shim and safety rods are sometimes combined in rods that have low rates of withdrawal but that can be rapidly inserted. This is usually done when the effect of burnup is to decrease reactivity. The rods are only partially inserted at the outset of operation, but the reactor can be quickly shut down by lowering them all the way into the core (scramming). As operation proceeds, the rods are moved farther out so that there is a greater shutdown reactivity margin.

The amount of shim control required can be reduced by the use of a burnable "poison." This is a neutron-absorbing material, such as boron or gadolinium, which will burn off faster than the fissile material does. At the beginning of operation, this controls the extra reactivity that has been built into the fuel to compensate for the amount of fuel consumed. At the end of an operating period, the absorber material will have been almost completely destroyed by neutron capture.

Structural components. These are the parts of a reactor system that hold the reactor together and permit it to function as a useful energy source. The most important structural component is usually the reactor vessel. In both the light-water reactor and the high-temperature gas-controlled reactor (HTGR), a pressure vessel is used so that the coolant can be contained and operated under conditions appropriate for power generation—namely, high temperature and pressure. Within the reactor vessel are structural grids for holding the reactor core and solid reflectors; coolant channels; control-rod guide channels; internal thermohydraulic components (*e.g.,* pumps or steam circulators) in some cases; instrument tubes; and parts of safety systems.

Pressure vessel

Coolant system. The function of a power reactor installation is to extract the heat of nuclear fission and convert it to useful power, generally electricity. The coolant system plays a pivotal role in performing this function. A coolant fluid enters the core at low temperature and leaves it at higher temperature. This higher temperature fluid is then directed to conventional thermodynamic components where the heat is converted into electrical power. In most light-water, heavy-water, and gas-cooled power reactors, the coolant is maintained at high pressure. Sodium and organic coolants operate at atmospheric pressure.

Research reactors have very simple heat removal systems in which coolant is run through the reactor and the heat that is removed is transferred to ambient air or to water without going through a power cycle. In research reactors of the lowest power running at only a few kilowatts, this may involve simple heat exchange to tap water or to a pool of water cooled with ambient air. During operation at higher power levels, the heat is usually removed by means of a small natural-draft cooling tower.

Containment system. Reactors are designed with the expectation that they will operate safely without releasing radioactivity to their surroundings. It is, however, recognized that accidents can occur. An approach using multiple barriers has been adopted to deal with such accidents. These barriers are, successively, the fuel cladding, primary vessel, and thick shielding. As a final barrier, the reactor is housed in a containment structure. This consists basically of the reactor building, which is designed and tested to prevent any radioactivity that escapes from the reactor from being released to the environment. As a consequence, the containment structure must be at least nominally airtight. In practice, it must be able to maintain its integrity under circumstances of a drastic nature, such as accidents in which most of the contents of the reactor core are released to the building. It has to withstand pressure buildups and damage from debris propelled by an explosion within the reactor, and it must pass a test to demonstrate that it will not leak more than a small fraction of its contents over a period of several days, even when its internal pressure is well above that of the surrounding air. The most common form of containment building is a cylindrical structure with a spherical dome, which is characteristic of LWR systems. This is much more typical of nuclear plants than the large cooling tower that is often used as a symbol for nuclear power. (It should be noted that cooling towers are found at large modern coal- and oil-fired power stations as well.)

Reactors other than those of the LWR type also have containment structures, but they vary in shape and construction. When it can be justified that major pressure buildups are not to be expected, the containment can be any form of airtight structure. In the United States, containment structures are required for all commercial power reactors and all high-power research reactors. In general, low-power research reactors are exempt, based on the common assumption that an accident in such systems will not lead to a widespread release of radioactivity. Reactors operated by the U.S. Department of Energy and by the armed services also are exempt, a matter which has caused considerable controversy. Some of these have containment structures, while others do not.

Relative integrity of containment

The concept of containment originated in the United States during the 1950s and has been generally accepted throughout much of the world. The Soviet bloc countries, however, did not concur with this view, and when containment was provided it was generally not up to Western standards. For example, Chernobyl Unit 4, which suffered a catastrophic explosive accident and fire in 1986, merely had an internal structure that could only withstand the loss of function of a single pressure tube. Though called containment, this was a misnomer by Western standards.

The most severe test of a containment system occurred during an accident in the United States in 1979 at Three Mile Island Unit 2, near Harrisburg, Pa. In this installation, a stoppage of core cooling resulted in the destruction, including partial melting, of the entire core and the release of a large part of its radioactivity to the enclosure around the reactor. In spite of a hydrogen deflagration that also occurred during the accident, the containment structure prevented all but a very small amount of radioactivity from entering the environment and must be credited with having prevented a major radioactive release and its consequences.

TYPES OF REACTORS

Most of the world's existing reactors are power reactors. There also are many research reactors, and the navies of many nations include submarines and surface ships driven by propulsion reactors. There are several types of power reactors, but only one, the light-water reactor, is widely

used. Accordingly, this variety is discussed in considerable detail here. Other significant types are briefly described, as are research and propulsion reactors. Some attention is also given to the prospective uses of reactors for space travel and for certain industrial purposes.

Power reactors. *Light-water reactor.* As noted above, LWRs are power reactors that are cooled and moderated with ordinary water. There are two basic types: the pressurized-water reactor (PWR) and the boiling-water reactor (BWR). In the first type, high-pressure, high-temperature water removes heat from the core and is then passed to a steam generator. Here the heat of the coolant is transferred to a stream of water in the generator (the secondary loop in Figure 41B), causing the water to boil and slightly superheat. The steam generated by this serves as the working fluid in a steam-turbine cycle (see *Steam turbines* above).

Pressurized-water reactor and boiling-water reactor

In a boiling-water reactor, water passing through the core is allowed to boil at intermediate pressure, and the steam from the reactor is used directly in the power cycle (see Figure 41A). Although the BWR seems simpler, the PWR has advantages with regard to fuel utilization and power density, and the two concepts have been economically competitive with each other since the 1960s. Both these light-water reactors are fueled with uranium dioxide pellets in zirconium alloy cladding (see above). The BWR fuel is slightly less enriched, but the PWR fuel produces more energy before being discharged, and so these two aspects balance each other out economically. Because the BWR operates at lower pressure, it has a thinner pressure vessel than the PWR; however, because its power density is somewhat lower, the BWR's vessel has a larger diameter for the same reactor power. The internal system of a BWR is more complex, since there are internal recirculation pumps and complex steam separation and drying equipment within its vessel. Though the internals of the PWR are simpler, a BWR power plant is smaller because it has no steam generators. In fact, the steam generators—there are usually four of them in a big PWR plant—are larger than the reactor vessel itself. The control rods of a typical PWR are inserted from the top (through the reactor head), while those of a BWR are inserted from the bottom.

Refueling operation

Light-water reactors are refueled by removing the reactor head—after lowering and unlatching the safety rods in the case of a PWR. This exposes the reactor to visual observation. The pressure vessel is filled to the top with water, and, since the core is near the bottom of the vessel, the water acts as a shield for this operation. Then, the fuel assemblies to be removed are lifted up into a shielded cask within which they are transferred to a storage pool for cooling while they are still highly radioactive. Many of the remaining assemblies are then shifted within the core, and finally fresh fuel is loaded into the empty fuel positions. The purpose of shifting fuel at the time of reload is to achieve an optimal reactivity and power distribution for the next cycle of operation. Reloading is a time-consuming operation. In principle, it could be accomplished in three weeks, but in practice the plant undergoes maintenance during reload, which can take considerably more time—up to a few months. Utilities schedule maintenance and reload during the spring and fall when electricity demand is lowest and the system usually has reserve capacity.

Storage pools for spent fuel

The discharged fuel stored in the storage pool is not only highly radioactive but also continues to produce energy. This energy is removed by natural circulation of the water in the storage pool. Originally it was expected that this spent fuel could be shipped out for reprocessing within two years, but this option is currently practiced only in France. In the United States, storage pools have continued to receive spent fuel, and some of the pools are filling up. Options available to nuclear plant operators are to store the spent fuel more densely than originally planned, to build new pools, or to store the oldest, no longer very hot fuel in above-ground silos (dry storage). Ultimately this fuel will be transferred to the U.S. Department of Energy for reprocessing or waste disposal or both, but this may not happen until the year 2003 or perhaps later if a viable disposal program is not established.

During the 1970s light-water reactors represented the cheapest source of new electricity in most parts of the world, and it still is economical in Japan, Korea, Taiwan, and France and many other European countries. In the United States, however, strict regulation of light-water reactors during the 1980s, coupled with a decrease in reactor research and development activity, have made

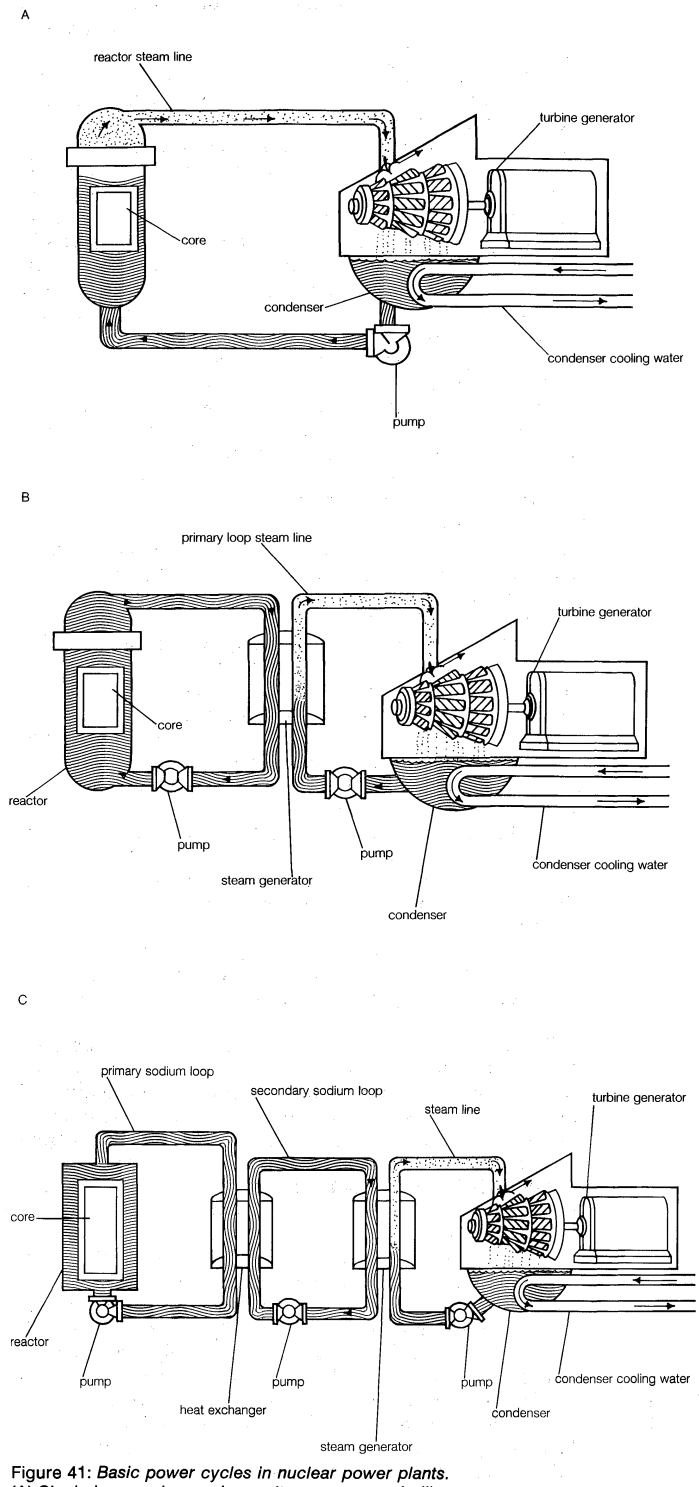


Figure 41: *Basic power cycles in nuclear power plants.* (A) Single-loop cycle; as shown, it represents a boiling-water reactor (BWR), but it could also represent a direct-cycle, high-temperature gas-cooled reactor (HTGR) if the word helium were substituted for the word steam. (B) Two-loop cycle; the primary loop depicted here could constitute a pressurized-water reactor (PWR), a CANDU pressurized heavy-water reactor (PHWR), or a helium HTGR. (C) Three-loop cycle; this is found only in sodium-cooled reactors where an intermediate loop of nonradioactive sodium is provided between the radioactive primary loop and the steam generator.

From the U.S. Department of Energy in R.A. Knief, *Nuclear Energy Technology* (1981), McGraw-Hill, Inc.

the competitive nature of new light-water reactor installations problematic. Plants that have been exceptionally well managed during construction and operation remain competitive; unfortunately, these are not the rule. New designs, developed abroad, may alter this situation, however.

Most recent light-water reactors have had electric capacity ratings of 1,000 megawatts or more. These are not very suitable for the utility industry, which has had only a slow growth in base-load demand since about 1975. Therefore, as of 1989, advanced light-water reactors in the 600-megawatt capacity range were also being considered.

High-temperature gas-cooled reactor. The HTGR, as mentioned above, is fueled with a mixture of graphite and fuel-bearing microspheres. There are two competitive designs of this reactor type: (1) a German system that uses spherical fuel elements of tennis-ball size loaded into a graphite silo and (2) an American version in which the fuel is loaded into precisely located graphite hexagonal prisms. In both variants, the coolant consists of helium pressurized to about 100 bars. In the German system the helium passes through interstices in the bed of the spherical fuel elements, while in the American system it passes through holes in the graphite prisms. Both are capable of operating at very high temperature, since graphite has an extremely high sublimation temperature and helium is completely inert chemically. The hot helium can be used directly as the working fluid in a high-temperature gas turbine, or its heat can be utilized to generate steam for a water cycle. Experimental prototypes of both the American and German designs have been built, but no commercial plants were on order as of 1989.

Sodium-cooled systems

Liquid-metal reactors. Sodium-cooled, fast-neutron-spectrum reactors received much attention during the 1960s and '70s when it appeared that their breeding capabilities would soon be needed to supply fissile material to a rapidly expanding nuclear industry. When it became clear in the 1980s that this was not a realistic expectation, enthusiasm slackened. The developmental work of the previous decades, however, resulted in the construction of a number of liquid-metal reactors around the world—in the United States, the Soviet Union, France, Britain, Japan, and Germany. Most liquid-metal reactors are fueled with uranium dioxide or mixed uranium–plutonium dioxides. In the United States, however, the greatest success has been with metal fuels. While some liquid-metal reactors are of the loop type, equipped with heat exchangers and pumps outside the primary reactor vessel, others are of the pool variety featuring a large volume of primary sodium in a pool that also contains the primary pumps and primary-to-secondary heat exchanger. In all types, the heat extracted from the core by primary sodium is transferred to a secondary, nonradioactive sodium loop, which serves as the heat source for a steam generator and turbine. The pool type seems to have some advantage in terms of safety in that the large volume of primary sodium heats up only slowly even if no power is extracted; thus, the reactor is effectively isolated from upsets in the balance of the plant. The reactor core in all such systems is a tightly packed bundle of fuel in steel cladding through which the sodium coolant flows to extract the heat. Most liquid-metal reactors are breeders or are capable of breeding, which is to say that they all produce more fissile material than they consume.

Natural uranium fuel

CANDU reactor. Canada focused its developmental efforts on reactors that would utilize abundant domestic natural uranium as fuel without having to resort to enrichment services that could be supplied only by other countries. The result of this policy was CANDU—the line of natural uranium-fueled reactors moderated and cooled by heavy water. A reactor of this kind consists of a tank, or calandria vessel, containing cold heavy water at normal pressure. The calandria is pierced by pressure tubes made of zirconium alloy, in which the natural uranium fuel is placed and the heavy water coolant is circulated. Power is obtained by transferring the heat from the exiting hot pressurized heavy water to a steam generator and then running the steam from the latter through a conventional turbine cycle. The fuel assembly of a CANDU reactor, which consists of a bundle of short zirconium alloy-clad tubes containing natural uranium dioxide pellets, can be changed while the system is running. A new assembly is simply pushed into one end of a pressure tube and the old one collected as it drops out at the other end. This feature has given the CANDU higher capacity factors than other reactor types. Several countries have purchased CANDU reactors for the same reason that they were developed by Canada—to be independent of imported enrichment services.

Advanced gas-cooled reactor. The advanced gas-cooled reactor (AGR) was developed in Britain as the successor to reactors of the Calder Hall class, which combined plutonium production and power generation. Calder Hall was the first nuclear station to feed an appreciable amount of power into a civilian network. It was fueled with slugs of natural uranium metal canned in aluminum, cooled with carbon dioxide, and employed a moderator consisting of a block of graphite pierced by fuel channels. In the advanced gas-cooled reactor, fuel pins clad in Zircaloy (trademark for alloys of zirconium having low percentages of chromium, nickel, iron, and tin) and loaded with 2-percent enriched uranium dioxide are placed into zirconium-alloy channels that pierce a graphite moderator block. The enriched fuel permits operation to economic levels of fuel burnup. A coolant of carbon dioxide transports heat to a steam generator, activating a steam-turbine cycle. Although a number of advanced gas-cooled reactors have been built in Britain, they have been less trouble-free and more costly than expected, and no new ones are planned.

Other power reactor types. A large variety of reactor types have been built and operated on an experimental basis. A few examples include organic liquid-cooled and -moderated reactors that can operate like a pressurized-water reactor without requiring high pressures in the primary circuit; sodium-cooled, graphite-moderated reactors; and heavy-water reactors built in a pressure-vessel design.

Research reactors. *Water-cooled, plate-fuel reactor.* This is the most common type of research reactor. It uses enriched uranium fuel in plate assemblies (see above) and is cooled with water. Water-cooled, plate-fuel reactors operate over a wide range of thermal power levels, from a few kilowatts to hundreds of megawatts. The systems with the lowest power ratings are usually operated at universities and used primarily for teaching, while those with the highest are used by major research laboratories chiefly for materials testing and research.

Pool-type reactors

A common form of the water-cooled, plate-fuel reactor is the pool reactor, in which the reactor core is positioned at the bottom of a large, deep pool of water. This has the advantage of simplifying both observation and the placement of channels from which beams of neutrons can be extracted. At lower thermal power levels, no pumping is required and the cooling water circulates by natural convection. A heat exchanger is usually located at the top of the pool, where the hottest water is stratified. At higher operating power levels, pumping becomes necessary to augment the natural circulation.

Most pool reactors use the water of the pool as a reflector (see above), but some have blocks of a solid moderator (canned graphite or beryllium metal) around the core that serves as an inner reflector. Graphite and beryllium create a large peak in slow neutron intensity a short distance from the core, which is an advantage when beams of slow neutrons are to be extracted or when such neutrons are used for irradiating materials.

Tank-type reactors

At higher power levels, it becomes more convenient to employ a tank-type reactor because it is simpler to control the flow path of pumped water in such a system. Low-power teaching reactors also are available in the tank form. The core and reflector arrangement in tank-type, plate-fuel research reactors is the same as in the pool-type systems and has the same variations; however, solid concrete shielding is employed around the sides instead of the water shield characteristic of the latter.

TRIGA reactors. The TRIGA system is an increasingly popular variety of research reactor. It is another tank-type, water-cooled system, but its fuel differs from that employed by the above-mentioned research reactors. The fuel

assembly of the TRIGA consists of zirconium-clad rods of mixed uranium and zirconium hydrides. The virtue of this fuel is that it exhibits an extremely large negative power-reactivity coefficient—so large that the reactor can be made strongly supercritical for an instant, causing its power to rise very rapidly, after which it quickly shuts itself down. The resulting power pulse is useful for a number of dynamic experiments. The total energy released in a pulse is not a problem, since the automatic shutdown occurs very quickly and the energy release is proportional to both peak power and pulse duration.

Other research reactors. As in the case of power reactors, a number of different reactor types have seen service as research reactors, and some are still in operation. The variety is so great as to defy cataloging. There have been homogeneous (fueled solution cores), fast, graphite-moderated, heavy-water-moderated, and beryllium-moderated reactors, as well as those adapted to use fuels left over from power reactor experiments. The design of research reactors is much more fluid and sensitive to a greater variety of special research demands than is design for other applications.

Ship propulsion reactors. The original, and still the major, naval application of nuclear energy is the propulsion of submarines. The chief advantage of using nuclear reactors for submarine propulsion is that they, unlike fossil-fuel combustion systems, require no air for power generation. Consequently, a nuclear-powered submarine can remain underwater indefinitely, whereas a conventional diesel-powered submarine must surface periodically to run its engines in air. Nuclear power confers a strategic advantage on naval surface vessels as well because it eliminates their dependence on refueling from vulnerable tankers.

Advantages of nuclear-powered naval craft

The design of U.S. naval nuclear power plants is classified for defense security purposes, and so only general information pertaining to them has been published. It is known that such power plants are fueled with highly enriched uranium and moderated and cooled with light water. The design of the first nuclear submarine power plant, that of the USS *Nautilus,* was heavily influenced by high-power research reactor design. Special features include the incorporation of a very large reactivity margin to accommodate long burnups without refueling and to permit restart after shutdown. For submarine use, the power plant also must be extremely quiet to avoid sonic detection. Various models have been developed to fit the specific requirements of different classes of submarines.

The nuclear power plants for U.S. aircraft carriers are believed to have been derived from the power plant designs for the largest submarines, but again the particulars of their design have not been published.

Besides the United States, Britain, France, and the Soviet Union have nuclear submarines. In each case, the design was developed in secret, but it is generally believed that they are all rather similar; the demands of the application usually lead to similar solutions. The Soviet Union also has a small fleet of nuclear-powered icebreakers, whose power plants are thought to be essentially the same as those in their earliest submarines. As with naval vessels, the ability to operate without refueling is an enormous advantage for Arctic icebreakers.

Nuclear-powered icebreakers

Prototypes of nuclear-powered commercial cargo ships were built and operated by the United States and West Germany but have now been decommissioned. These vessels did not operate very economically, and opposition to their docking in a number of major ports also was a factor in their decommissioning. The prototypes were powered by reactors of the pressurized-water type.

Production reactors. The very first nuclear reactors were built for the express purpose of manufacturing plutonium for nuclear weapons, and the euphemism of calling them production reactors has persisted to this day. At present, most of the material produced by such systems is tritium (^{3}H, or T), the fuel for hydrogen bombs. Plutonium has a long half-life, and so countries with arsenals of nuclear weapons using plutonium as fissile material generally have more than they expect to need. On the other hand, tritium has a half-life of only about 12 years; thus stocks of this radioactive hydrogen isotope have to be continuously replenished. The United States, for example, operates several reactors moderated and cooled by heavy water that produce tritium at the Savannah River facility in South Carolina.

The plutonium isotope that is most desirable for sophisticated nuclear weapons is plutonium-239. If plutonium-239 is left in a reactor for a long time after production, plutonium-240 builds up as an undesirable contaminant. Accordingly, a major feature of a production reactor is its capability for quick throughput of fuel at a low energy-production level. Any reactor that can be operated this way is a potential production reactor.

The world's first plutonium production reactors, built by the United States at Hanford, Wash., were fueled with natural uranium, moderated by graphite, and cooled by light water. It is believed that the early Soviet production reactors were the same sort, and the French and British versions differed only in that they were cooled with gas. As was noted above, the first significant power reactor, the Calder Hall reactor, was actually a dual-purpose production reactor.

Specialized reactors. Nuclear reactors have been developed to provide electric power and steam heat in far-removed, isolated areas. The Soviet Union, for instance, has installed smaller power reactors specially designed to supply both electricity and steam for heating to accommodate the needs of a number of remote Arctic communities. Independent developmental work on small automatically operated reactors with similar capabilities has been undertaken by Sweden and Canada.

Reactors have been developed to supply power and propulsion in space. The Soviet Union has been deploying small intermediate reactors in satellites for powering equipment and telemetry since the 1970s, but this policy has drawn criticism because at least one reactor-powered spacecraft has reentered the atmosphere and deposited radioactive debris in Canada. Developmental activity in the United States has been directed largely toward reactor applications for the Strategic Defense Initiative (SDI) and for such deep-space missions as manned exploration of other planets or the establishment of a permanent lunar base. Reactors for these applications would necessarily be high-temperature systems based on either the HTGR or the LMR design but that would use enriched fuel. A power cycle in space must be run at a very high temperature to minimize the size of the radiator from which heat is to be rejected. A reactor for space applications also has to be compact so that it can be shielded with a minimum amount of material.

Nuclear-powered spacecraft

Small pressurized-water reactors have been used in the past to provide power for remote bases in Greenland and Antarctica. Though they have been replaced with oil-fired power plants, it still appears feasible to employ nuclear power for such applications or even for more exotic ones, such as supplying power to permanent undersea camps.

Finally, concepts have been developed, notably in West Germany, for employing HTGR systems as sources of high-temperature heat for chemical process industries. An idea that has drawn particular attention involves the use of reactor-generated heat at the mouth of a coal mine to convert the coal into clean gas for delivery by pipeline. Such processes remain economically unattractive at present but may ultimately became feasible as natural sources of fluid fuels are exhausted.

REACTOR SAFETY

Nuclear reactors contain very large amounts of radioactive isotopes—mostly fission products but also such heavy elements as plutonium. If this radioactivity were to escape the reactor, its effects on the people in the vicinity would be severe. The deleterious effects of exposure to high levels of ionizing radiation would include increased rates of cancer and genetic defects, an increased number of developmental abnormalities in children exposed in the womb, and even death within a period of several days to months when irradiation is extreme (see RADIATION: *Major types of radiation injury*). For this reason, a major consideration in reactor design is ensuring that a significant release of radioactivity does not occur. This is ac-

Safeguarding against radioactive contamination of the environment

complished by a combination of preventive measures and mitigating measures. Preventive measures are those that are taken to avoid accidents, and mitigating measures are those that decrease the adverse consequences. Essentially, preventive measures are the set of design and operating rules that are intended to make certain that the reactor is operated safely, while mitigating measures are systems and structures that prevent such accidents as do occur from proceeding to a catastrophic conclusion. Among the most well-known preventive measures are the reports and inspections for double-checking that a plant is properly constructed; rules of operation; and qualification tests for operating personnel to ensure that they know their jobs. The mitigating measures include safety rod systems for quickly shutting down a reactor to prevent a runaway chain reaction; emergency cooling systems for removing the heat of radioactive decay in the event that normal cooling capability is lost; and the containment structure for confining any radioactivity that might escape the primary reactor system. An extreme mitigating measure is the exercising of plans to evacuate personnel who might otherwise be heavily exposed in a reactor installation.

Preventive measures. Since no human activity can be shown to be absolutely safe, all these measures cannot reduce the risks to zero, but it is the aim of the rules and safety systems to minimize the risk to the point where a reasonable individual would conclude they are trivial. What this *de minimis* risk value is, and whether it has been achieved by the nuclear industry, is a subject of bitter controversy, but it is generally accepted that independent regulatory agencies—the United States Nuclear Regulatory Commission (NRC) and similar agencies around the world—are the proper judges of such matters.

To help evaluate the risks from nuclear power plants, the U.S. Atomic Energy Commission (AEC) authorized a major safety study in 1972 (the AEC was disbanded in 1974 and its functions have been assumed by the NRC). The study was conducted with major assistance from a number of laboratories, and it involved the application

Probabilistic risk assessment

of probabilistic risk assessment (PRA) techniques for the first time on a system as complex as a large nuclear power reactor. This work resulted in the publication in 1975 of a report titled *Reactor Safety Study,* also known as WASH-1400. The most useful aspect of the study was its delineation of components and accident sequences (scenarios) that were determined to be the most significant contributors to severe accidents.

The *Reactor Safety Study* concluded that the risks of an accident that would injure a large number of people were extremely low for the light-water reactor systems analyzed. This conclusion, however, was subject to very large quantitative uncertainties and was challenged.

One basic problem with probabilistic risk assessment is that it cannot easily be confirmed by experience when the level of risk has been reduced to low values. That is to say, if probabilistic risk assessment predicts that a reactor is subject to, say, one failure in 10,000 years, there is no way to prove that statement with only a few, or even with 10,000, years of experience. Thus, the results of the *Reactor Safety Study* as to risk levels were not confirmable.

There matters stood until 1979, when Three Mile Island Unit 2 suffered a severe accident. Through a combination of operator errors, coupled with the failure of an important valve to operate correctly, cooling water to the core was lost, parts of the core were melted and the rest of it destroyed, and a large quantity of fission products was released from the primary reactor system to the interior of the containment structure. The containment vessel of the reactor building fulfilled its function, and only a small amount of radioactivity was released, demonstrating the wisdom of having this component. Still, a severe accident had occurred.

Many investigations of the Three Mile Island accident followed. Recommendations differed among them, but a common thread was that the human element was a much more important factor in safe operation than had been hitherto recognized. The human element pertained not only to the operating staff but also to the managements of nuclear plants and even to the NRC itself. Following the accident, therefore, many changes in operator training and in technical and inspectorate staffing were implemented, just as a number of hardware enhancements were introduced. It is generally believed that these changes have been effective in reducing the likelihood of the occurrence of accidents as severe as that at Three Mile Island. As a side issue to this, however, the operating costs of nuclear power plants have escalated sharply as more and more highly trained people have been added to the operating staffs.

One area where probabilistic risk assessment has proven useful is with regard to the licensing of new plants, either light-water reactor installations or those of less common reactor types. PRA has the virtue of comparing systems fairly reliably. With better computer hardware and software than were available in 1975, it has become feasible to do PRA analyses of individual plants and compare them. A standard protocol for the NRC in licensing new, and particularly new types of, plants has therefore been that they must demonstrate lower risks than light-water reactors, which have been accepted as the norm.

The significance of the human element, particularly as it relates to plant management and high-level regulatory decision making, was borne out again by the Chernobyl catastrophe of 1986. One of the four reactors in a nuclear power station about 100 kilometres north of Kiev exploded and caught fire as the result of an ill-conceived experiment (a test to see how long the steam turbines would run while coasting to a stop if the reactor would be abruptly shut down). Before the situation had been brought under control, 31 people had died (two from the blast and 29 from radiation exposure), an estimated 25 percent of the radioactive contents of the reactor had been released in a high cloud plume, 135,000 people had to be evacuated, and a large area surrounding the plant received fallout so great that it could not be farmed or pastured. Significant radiation was detected as far north as Scandinavia and as far west as Switzerland. It has been estimated that between 4,000 and 40,000 cases of cancer would ultimately result from this accident (besides the initial several hundred victims), mostly within the Soviet Union but some in areas far removed from there. Investigation of the accident placed the largest blame, as with the Three Mile Island mishap, on poor management both at the plant and within the government bureaucracy.

The Chernobyl disaster of 1986

Because all such nuclear plant accidents have basically resulted from human failings rather than from some intrinsic factor, most experts believe that nuclear energy can be a safe source of power. A review of the overall performance record shows that there had been, as of 1989, several thousand "reactor-years" of safe power-reactor operation in the Western world, with health effects less damaging than those associated with the extraction of an equal amount of power from coal. Incorporating the lessons learned from past accidents should certainly make future operations safer. There is, however, a condition on the conclusion that nuclear power is by and large a safe form of power. The facilities for generating this power must be designed, built, and operated to high standards by knowledgeable, well-trained professionals; and a regulatory mechanism capable of enforcing these standards must be in place.

Mitigating measures. Two of the principal safety measures, the safety rods and the containment structure, have already been described. Other major safety systems are the emergency core cooling system, which makes it possible to cool the reactor if normal cooling is disrupted, and the emergency power system, which is designed to supply electrical power in case the normal supply is disrupted so that detectors and vital pumps and valves can continue to be operated. An important part of the safety system is the strict adherence to design rules, some of which have been mentioned—namely, the reactor should have a negative power-reactivity coefficient; the safety rods must be injectable under all circumstances; and no single regulating rod should be able to add substantial reactivity rapidly. Another important design rule is that the structural materials used in the reactor must retain acceptable physical properties over their expected service life. Finally, construction is to be covered by stringent quality assurance

rules, and both design and construction must be in accordance with standards set by major engineering societies and accepted by the NRC.

Most probable accidents and risks

According to probabilistic risk assessment studies, three kinds of events are most responsible for the risks associated with light-water reactors—namely, station blackout, transient without scram, and loss of cooling. The nature of each of these mishaps is delineated, as are the proposed countermeasures and the anticipated risks.

In station blackout, a failure in the power line to which the station is connected is postulated. The proposed emergency defense is a secondary electrical system, typically a combination of diesel generators big enough to drive the pumps and a battery supply sufficient to run the instruments. The risk would be that of the emergency generators not accepting load when they are started up. In transient without scram, the event is insertion of reactivity, for example, by an unchecked withdrawal of shim rods. The protective response is the rapid and automatic insertion of the safety rods. The risk would be the safety rods not functioning properly. In loss of cooling, the event is a failure of the normal cooling system to operate, either because of a break in a coolant line or because of an operator error. The emergency response is activation of the emergency core cooling system, and the risk would be that the system fails to operate. The ultimate event in the chain that led to the Three Mile Island accident was loss of emergency cooling by operator action owing to a misinterpretation of what sort of accident was occurring. In all these cases, proper operator action as well as proper functioning of the appropriate backup system are important aspects of emergency response. A final backup capability that is coming into play is the use of computers in an advisory mode to help the operator understand what is happening and suggest proper responses.

Different reactor types pose different types of risk. For example, neither the pool-type liquid-metal reactor nor the high-temperature gas-cooled reactor are at major risk with regard to loss of cooling and perhaps not with regard to station blackout. However, the LMR, and perhaps the HTGR, are at some risk from events that might cause air or water to enter the coolant system. The hazard is that reactor materials, sodium or graphite, could chemically react with air and water. The hazard is greater with sodium in the LMR than it is with graphite in the HTGR.

Another type of risk arises from external events, such as the possibility that earthquakes might initiate one or another major accident. The earthquake risk is minimized by building plants away from faults and by making use of earthquake-resistant mechanical design and construction.

NUCLEAR FUEL CYCLE

No discussion of nuclear power can be complete without a brief exposition of the nuclear fuel cycle. The whole point of a reactor is, after all, to cause fission in nuclear fuel. Moreover, it has turned out that low cost of fueling is the chief reason for the economic competitiveness of nuclear power.

Principal steps of the fuel cycle

The principal steps of the fuel cycle are uranium mining and extraction from its ore (milling), uranium enrichment, fuel fabrication, loading and irradiation in the reactor (fuel management), unloading and cooling, reprocessing, waste packaging, and waste disposal (see Figure 42).

Uranium mining. Uranium is mined from ores whose uranium content is on the order of 0.1 percent (one part per thousand). Most ore deposits are at or near the surface, and whether they are mined by open-pit or deep-mining techniques depends on the depth of the deposit and whether it slopes downward. The ore is crushed and the uranium chemically extracted from it at the mouth of the mine. The residue remains radioactive as it contains long-lived radioactive daughter nuclei of uranium and has to be carefully managed to minimize the release of radioactive contaminants into the environment. The uranium concentrate, which consists of uranium compounds (typically 75 to 95 percent), is shipped to a chemical plant for further purification and chemical conversion.

Enrichment. There are several possible enrichment methods, but the only two that are used on a large scale

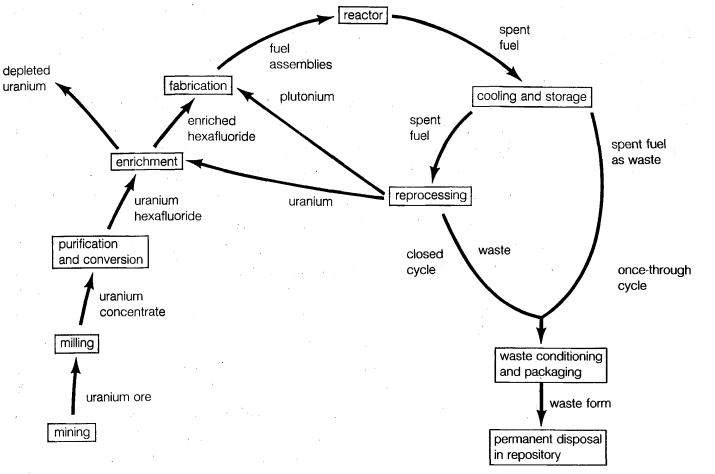


Figure 42: Light-water reactor fuel cycle.

are gaseous diffusion and gas centrifuging. In gaseous diffusion, natural uranium in the form of uranium hexafluoride gas (UF_6), a product of chemical conversion, is allowed to seep through a porous barrier. The molecules of $^{235}UF_6$ penetrate the barrier slightly faster than those of $^{238}UF_6$. Since the percentage of ^{235}U increases by only a very small amount after traversal of the barrier, the process must be repeated over and over in a large number of stages to obtain the desired amount of enrichment.

Gaseous diffusion

In gas centrifuging, the uranium hexafluoride gas is fed into a high-speed centrifuge. The lighter species of this mixture of gaseous molecules including ^{235}U tend to concentrate away from the wall, while the heavier ones accumulate along the wall. The degree of enrichment per stage in a centrifuge is greater than that obtained in a gaseous diffusion chamber, but the centrifuge is a more expensive piece of equipment.

Fabrication. This step involves the conversion of the suitably enriched product material to the chemical form desired for reactor fuel. As of the late 1980s the only fuel fabricated on a large scale was that for light-water reactors.

The chemical form prepared for the light-water reactor is uranium dioxide. Produced in the form of a ceramic powder, this compound is ground into a very fine flour and inserted into a die, where it is pressed into a pellet shape. Next the pellet is sintered in a furnace at 1,500–1,800° C. This sintering, similar to the firing of other ceramic ware, produces a dense ceramic pellet. Such pellets are loaded into prefabricated zirconium alloy cladding tubes, which are then filled with an inert gas and welded shut. These tubes, or pins, are bundled together with proper spacing assured by top and bottom grid plates through which the ends of the pins pass. Together with other necessary hardware, the bundle constitutes a fuel assembly (Figure 43).

Fuel management. Fuel is loaded into a reactor in a careful pattern so as to obtain the most energy production from it before it becomes no longer usable. Fresh fuel is more reactive than old fuel, and this reactivity is used to keep the reactor critical. Typically, a reactor is fueled in cycles, each cycle lasting one to two years, and a fuel batch is kept in the reactor for three or four cycles. At the end of each cycle, the oldest fuel is removed and fresh fuel loaded. The partially burned fuel that remains, however, is shuffled before the fresh fuel is installed. The objective of this procedure is to achieve a loading of maximum reactivity while keeping the power distribution among the different fuel assemblies within technical specifications.

Factors that limit energy production

Fuel burnup—that is, energy production—is limited by two factors. After significant burnup has occurred, the physical properties of the fuel become degraded and it is not prudent to continue to keep it in the reactor. Also, after some burnup, the old fuel no longer contributes useful reactivity to the reactor. The fuel design, including its initial enrichment, is such that these two limits are made to approximately coincide.

Unloading and cooling. Spent reactor fuel is extremely radioactive, and its radioactivity also makes it a source of

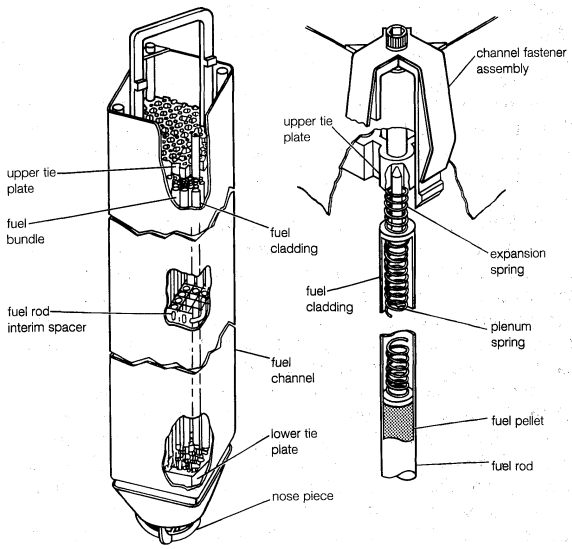


Figure 43: Fuel assembly.
From *General Description of a Boiling Water Reactor*, General Electric Co. (1974)

heat. When the spent fuel is removed from the reactor, it must continue to be both shielded and cooled. This is accomplished by placing the spent fuel in a water storage pool located next to the reactor. The water in the pool contains a large amount of dissolved boric acid, which is a heavy absorber of neutrons; this assures that the fuel assemblies in the pool will not go critical. (Pool water is also a common source of emergency cooling water for the reactor.) Pools vary in size; the older ones are only able to accommodate about 10 years worth of spent fuel. As the pools fill up, more spent fuel storage is needed. As noted earlier, additional storage space can be gained by loading spent fuel into the pool more densely than originally planned, by building a new pool, or by removing the oldest fuel assemblies from the existing pool and storing them in air-cooled concrete and steel silos located above ground. This last method becomes feasible after fuel has been stored for two or three years because radioactivity and heat generation decrease rapidly over this period. Dense storage in existing pools and silo storage both seem to be less expensive than building a new pool.

Reprocessing. Both the converted plutonium and residual uranium-235 in spent fuel can be recycled. Such materials can be recovered by chemically reprocessing the fuel. Equally as significant, reprocessing can reduce the volume and radioactivity of the waste material, which must ultimately be eliminated by some method of permanent disposal. Until 1975 it was generally assumed that after two to five years spent fuel would be delivered to a reprocessing plant. By that time, however, the cost of reprocessing had escalated to a point where its economics became questionable. Also, during the period 1976–81, it was U.S. policy, by presidential directive, not to reprocess. The directive has since been rescinded, but reprocessing is still not done commercially in the United States.

Policy and institutional arrangements are different in France and Britain. Commercial reprocessing plants exist in both countries and are processing spent fuel not only from nuclear plants in the host countries but also from those in others. The reprocessed plutonium can be used not only as fuel for planned future liquid-metal reactors but also to help fuel existing light-water reactors. In the latter application, the plutonium is utilized in mixed oxide form—a combination of uranium and plutonium dioxides having 3 to 6 percent plutonium.

Methodology

Reprocessing is accomplished by dissolving the spent fuel in nitric acid and contacting the acid solution with oil in which tributyl phosphate (TBP) is dissolved. TBP is a complexing agent for uranium and plutonium, forming compounds with them that bring them into the oil solution. A physical separation of the (immiscible) oil and acid serves to remove the desired products from the nitric acid solution, which still contains all the fission products. The uranium and plutonium can then be washed out of the TBP back into a water solution and separated from each other to the degree desired by means of various techniques. Thus, reprocessing produces three product streams: (1) a purified uranium product, (2) a plutonium product that may be either pure or mixed with uranium, and (3) a waste stream of fission products dissolved in nitric acid.

Waste conditioning. In the absence of reprocessing, the spent fuel is considered to be waste and must be prepared for disposal. This operation is to be performed in a separate facility, for which the Department of Energy has responsibility in the United States. As of 1998, the department is to begin receiving spent fuel from utilities largely on an "oldest-fuel-first" schedule. After brief storage, the fuel pins would be removed from their assemblies. End pieces that contain no fuel would be removed and the pins repacked into a dense lattice emplaced in a corrosion-resistant steel canister. A cover would be welded on and the canister covered with an overpack. This would represent the basic waste form for spent-fuel disposal.

Some waste exists in the form of the fission-product solution that arises from reprocessing. Reprocessed fuel from production reactors also generates this type of waste. The waste solution is completely evaporated, leaving behind the fission products in the solid residue, which is heated until all the constituent nitrate salts are converted to oxides. These oxides are then put into a glass-forming oven and mixed with materials that will produce a borosilicate glass. The fission-product oxides dissolve in the glass as it forms. The glass melt is subsequently poured into a steel canister, 200–400 millimetres in diameter and about one metre high, where it solidifies into a solid glass block. Once covered with an overpack of bentonite clay, the solid canister-like block is ready for disposal.

Waste solution from reprocessing

The glassmaking process for waste conditioning described here is operational on an industrial scale in France and has been tested in many other countries, including the United States.

Waste disposal. *Proposed method.* The waste disposal method currently being planned by all countries with nuclear power plants is called geologic disposal. This means that all conditioned nuclear wastes are to be deposited in mined cavities deep underground. Shafts are to be sunk into a solid rock stratum, with tunnel corridors extending horizontally from the central shaft region and tunnel "rooms" laterally from the corridors. The waste would be emplaced (probably by remotely controlled or robotic devices) in holes drilled into the floors of these rooms, after which the boreholes would be sealed and the rooms and corridors backfilled. When the entire operation is completed (perhaps after about 30 years of operation), the shafts too would be backfilled and sealed.

Geologic disposal method

Risks of nuclear waste disposal. Nuclear waste disposal is viewed quite differently by those who have studied it carefully and by the general public. When a holistic view of the process is taken, the risks seem extremely small, but it is one of the most feared aspects of the nuclear fuel cycle. A great deal of suspicion about the process arises from the numerous incidents of mismanagement of other types of waste, and these fears are encouraged by antinuclear activists. A number of basic observations on the process of geologic disposal point to the difficulty of resolving differences that are founded on perceptual discrepancies.

Nuclear waste retains its very intense level of radioactivity for several hundred years, but after 1,000 years have passed the remaining radioactivity, while persistent, is at a level comparable to, but greater than, that of a body of natural uranium ore. This separates the safety problem into two time periods: a first millennium during which it is crucial to ensure tight retention of the wastes in the repository, and a subsequent period during which it is only necessary to ensure that any release that occurs is small and slow.

There is general consensus that only impingement of groundwater and subsequent corrosion of the waste canisters, followed by dissolution of the waste, provides a route for the emergence of the waste in the surface environment. Water migrates slowly in most rock formations. Contrary to the popular belief that any dissolution of the waste and

discharge of the resulting solution to the environment will quickly lead to high-level contamination, only a low level is projected, even in worst-case scenarios.

Migration of radioactive species that has been observed at shallow burial sites for low-level radioactive waste is not an indication that similar migration can be expected in a deep underground repository. In addition to the near insolubility of the waste material, waste form engineering, particularly of corrosion-resistant containers, provides extra protection against such dispersal. Moreover, most of the dispersal problem in shallow disposal sites is caused by biochemical products that do not exist in deep formations; water found at depth is sterile.

Site selection

Finally, a great deal of care is to be expended in selecting the site of the repository. Site selection is probably the biggest problem, both politically and technically. Various conditions are mandatory: the repository must not be near a populated area, and the rock stratum selected must be deep (300 metres or more), dry, and naturally sealed from aquifers; and the water table should discharge only slowly into surface waters. Furthermore, the site must be in a tectonically inactive zone so that earthquakes will not break that seal.

The risk of high-level waste burial is almost certainly smaller than the risks of reactor accidents and even than the risks arising from improperly managed mine tailings. Nonetheless, the siting of a repository must be handled with political sensitivity, and the confirmation of acceptable hydrologic and geologic conditions must have a high degree of validity. There are many acceptable sites in principle, but confirming acceptability for any one of them is a large and expensive technical undertaking.

HISTORY OF REACTOR DEVELOPMENT

Soon after the discovery of nuclear fission was announced in 1939, it was also determined that the fissile isotope involved in the reaction was uranium-238 and that neutrons were emitted in the process. Newspaper articles reporting the discovery mentioned the possibility that a fission chain reaction could be exploited as a source of power. World War II, however, began in Europe in September of 1939, and physicists in fission research turned their thoughts to using the chain reaction in a bomb. It was quickly recognized that a high concentration of fissile material would be needed to accomplish this.

Inasmuch as fission had been first discovered in Germany, there was great fear, particularly among refugee physicists from Europe who had fled to America, France, and Britain, that Nazi Germany might develop just such a bomb. As a result, these three countries began working toward the development of atomic bombs, which at that point was still speculation. The most successful program was established in the United States, where President Franklin D. Roosevelt was persuaded by a letter from Albert Einstein to initiate a secret project devoted to this purpose. In early 1940 the U.S. government made funds available for research that eventually evolved into the Manhattan Project. After the fall of France to the German armies (1940), leading French researchers escaped to England and joined the ongoing British project. After the entry of the United States into the war in 1941, the British effort was transferred to the safer confines of North America. Though the British group participated in American research, it was chiefly concerned with initiating a research program in Canada.

The Manhattan Project

The Manhattan Project included work on uranium enrichment to procure uranium-235 in high concentrations and also research on reactor development. The goal was twofold: to learn more about the chain reaction for bomb design and to develop a way of producing a new element, plutonium, which was expected to be fissile and could be isolated from uranium chemically.

Reactor development was placed under the supervision of the leading experimental nuclear physicist of the era, Enrico Fermi. Fermi's project, begun at Columbia University and first demonstrated at the University of Chicago, centred on the design of a graphite-moderated reactor. It was soon recognized that heavy water was a better moderator and would be more easily used in a reactor, and this possibility was assigned to the Canadian research team since heavy-water production facilities already existed in Canada. Fermi's work led the way, and on Dec. 2, 1942, he reported having produced the first self-sustaining chain reaction. His reactor, later called Chicago Pile No. 1 (CP-1), was made of pure graphite in which uranium metal slugs were loaded toward the centre with uranium oxide lumps around the edges. This device had no cooling system, as it was expected to be operated for purely experimental purposes at very low power. CP-1 was subsequently dismantled and reconstructed at a new laboratory site in the suburbs of Chicago, the original headquarters of what is now Argonne National Laboratory. The device saw continued service as a research reactor until it was finally decommissioned in 1953.

Chicago Pile No. 1

On the heels of the successful CP-1 experiment, plans were quickly drafted for the construction of the first production reactors. These were the early Hanford reactors, which were graphite-moderated, natural uranium-fueled, water-cooled devices. As a backup project, a production reactor of air-cooled design was built at Oak Ridge, Tenn.; when the Hanford facilities proved successful, this reactor was completed to serve as the X-10 reactor at what is now Oak Ridge National Laboratory. Shortly after the end of World War II, the Canadian project succeeded in building a zero-power, natural uranium-fueled research reactor, the so-called ZEEP (Zero-Energy Experimental Pile). The first enriched-fuel research reactor was completed at Los Alamos, N.M., at about this time as enriched uranium-235 became available for research purposes (see Table 3). In 1947 a 100-kilowatt reactor with a graphite moderator and uranium metal fuel was constructed in England, and a similar one was built in France the following year.

Table 3: Notable Early Nuclear Reactors

name	location	power output*	distinction	start-up
CP-1 (Chicago Pile No. 1)	Chicago	low	first reactor	1942
ORNL Graphite, or Oak Ridge Graphite Reactor (X = 10)	Oak Ridge, Tenn.	3.8 MW	first megawatt-range reactor	1943
Y-Boiler (LOPO)	Los Alamos, N.M.	low	first enriched-fuel reactor	1944
CP-3 (Chicago Pile No. 3)	Chicago	300 kW	first heavy-water reactor	1944
ZEEP (Zero-Energy Experimental Pile)	Chalk River, Ont.	low	first Canadian reactor	1945
Hanford	Richland, Wash.	>100 MW	first high-power reactor	1945
Clementine	Los Alamos, N.M.	25 kW	first fast-neutron spectrum reactor	1946
NRX	Chalk River, Ont.	42 MW	first high-flux research reactor	1947
GLEEP	Harwell, Eng.	low	first British reactor	1947
ZOE (EL-1)	Châtillon, Fr.	150 kW	first French reactor	1948
EBR-1 (Experimental Breeder Reactor No. 1)	Idaho Falls, Idaho	1.4 MW	first breeder and first reactor system to produce electricity	1951
LITR (Low-Intensity Test Reactor)	Oak Ridge, Tenn.	3 MW	first plate-fuel reactor	1950
JEEP-1	Kjeller, Nor.	350 kW	first international reactor (Norway–Netherlands)	1951
STR (Submarine Thermal Reactor)	Idaho Falls, Idaho		submarine reactor prototype	1953
BORAX-III	Idaho Falls, Idaho	3.5 MW(e)	first U.S. reactor capable of significant electric power generation	1955
Calder Hall A	Calder Hall, Eng.	20 MW(e)	world's first reactor for large-scale commercial power production	1956

*Power output is thermal except where noted as MW(e), signifying electrical.

In 1953 President Dwight D. Eisenhower of the United States announced the Atoms for Peace program. This program established the groundwork for a formal U.S. nuclear power program and expedited international cooperation on nuclear power.

Development of commercial power reactors

The earliest U.S. nuclear power project had been started in 1946 at Oak Ridge, but the program was abandoned in 1948, with most of its personnel being transferred to the naval reactor program that produced the first nuclear-powered submarine, the *Nautilus*. After 1953 the U.S. nuclear power program was devoted to the development of several reactor types, of which three ultimately proved to be successful in the sense that they remain as commercial reactor types or as systems scheduled for future commercial use. These three were the fast breeder reactor (now called LMR); the pressurized-water reactor; and the boiling-water reactor. The first LMR was the Experimental Breeder Reactor, EBR-I, which was designed at Argonne National Laboratory and constructed at what is now the Idaho National Engineering Laboratory near Idaho Falls, Idaho. EBR-I was an early experiment to demonstrate breeding, and in 1951 it produced electricity from nuclear heat for the first time. As part of the U.S. nuclear power program, a much larger experimental breeder, EBR-II was developed and put into service (with power generation) in 1963. The principle of the boiling-water reactor was first demonstrated in a research reactor in Oak Ridge, but development of this reactor type was also assigned to Argonne, which built a series of experimental systems designated BORAX in Idaho. One of these, BORAX-III, became the first U.S. reactor to put power into a utility line on a continuous basis. A true prototype, the Experimental Boiling Water Reactor, was commissioned in 1957. The principle of the pressurized-water reactor had already been demonstrated in naval reactors, and the Bettis Atomic Power Laboratory of the naval reactor program was assigned to build a civilian prototype at Shippingport, Pa. This reactor, the largest of the power-reactor prototypes, is often hailed as the first commercial-scale reactor in the United States.

During the late 1950s and early 1960s a number of true commercial prototype nuclear power plants were built. Of these, the most successful was the light-water reactor system, although the advanced gas-cooled type remained the British standard for many years and the CANDU system prevailed in Canada. From the mid-1960s, larger units were ordered in the expectation of an ever-increasing commercial utilization of nuclear power, and by the early 1970s nuclear plant orders were coming in at such a rapid pace that the unit sizes were increased so as to reduce the number of separate projects that each vendor would have to staff for. By the later years of the decade, however, the surfeit of orders in the United States was followed by a large number of project cancellations. This phenomenon was the result of a sharp decrease over what had been projected as the rate of increase in base-load electricity demand for which the large nuclear plants were designed.

Escalating cost of building new nuclear plants

The new plants were not needed. Moreover, the cost of new nuclear plants had begun to escalate to the point where their economics became questionable. Public fears of nuclear power, stimulated by the Three Mile Island accident, also were a factor.

Similar scenarios have slowed the deployment of nuclear power in several countries besides the United States. On the other hand, France, Japan, South Korea, and Taiwan, which all have few alternative fuel resources, have continued building up their nuclear power capacity. (B.I.S.)

Electric generators and electric motors

A machine that converts mechanical energy into electrical energy is known as an electric generator. One that converts electrical energy into mechanical energy is an electric motor. Actually, the same machine can have energy flow in either direction. The same basic principles apply for both generators and motors. The designation as a generator or motor depends on the intended application.

Major uses

The major use of generators is to produce electrical power for distribution on transmission lines to domestic, commercial, and industrial customers. Generators also produce the electrical power required for automobiles, aircraft, ships, and trains.

Electric motors drive all sorts of mechanical devices. Typical examples are fans and windshield wipers on automobiles, elevators and air conditioners in office buildings, mixers and record players in homes, subway trains in cities, pumps in pipelines, and robots in industry.

Most electric machines rotate, often at high speed. Some, however, produce linear motion such as is required in rapid-transit vehicles.

BASIC PRINCIPLES OF OPERATION

Energy conversion by means of a magnetic field

Most electric machines convert energy by use of a magnetic field that allows force to be transmitted from a stationary to a moving part without physical connection. There are two basic principles exploited in generator and motor operation. The first, originally discovered by the French physicist André-Marie Ampère, states that an electrical conductor carrying a current at right angles to a magnetic field will experience a force at right angles to both the field and the current. The second principle, formulated on the basis of observations made by the English scientist Michael Faraday, states that a potential difference, or voltage, will be established between the ends of an electrical conductor that moves across or perpendicular to a magnetic field. These principles apply for a moving conductor in a stationary magnetic field. They apply equally for a stationary conductor with a moving magnetic field. The various configurations of electric machines consist of means of creating the magnetic field and placing current-carrying conductors in it in such a way as to produce force and voltage.

Elementary generators. These principles are demonstrated in the arrangement shown in Figure 44, where a loop of a conductor is rotated in a magnetic field. This field is created by the use of permanent magnets on each side, directing a horizontal field across a pair of air gaps. A central iron core and an outer iron yoke are used to provide an easy path for the magnetic field to close on itself and thus concentrate the field into the air gaps. Suppose the loop is rotated counterclockwise. In the left-hand section of the loop traveling downward across the field, positive electric charges will be forced toward the observer and negative charges will be forced away. On the right-hand section of the loop, the upward motion across the field forces negative charge toward the observer. The result is the establishment of a potential difference, or voltage, between the two terminals of the loop. This potential difference is proportional to the rate per second at which the magnetic field is being crossed by the two sides of the loop; that is to say, it depends on the density of the field, the length of conductor perpendicular to the field, and the velocity of the conductor perpendicular to the field.

If an electrical load such as a resistor is now connected

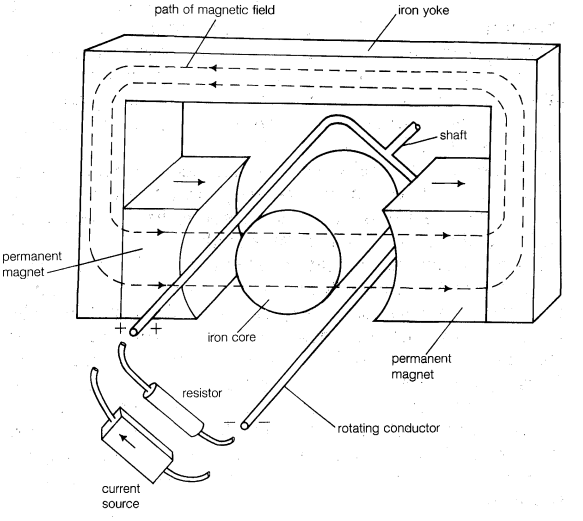


Figure 44: Elementary electric motor.

between the two terminals of the loop, an electric current will flow out of the positive terminal and into the negative terminal. This current will then interact with the magnetic field, resulting in an upward force on the left-hand conductor and a downward force on the right-hand one—*i.e.,* a force perpendicular to both the field and the current direction. To overcome this force and make the loop turn, mechanical torque must be applied to the shaft of the loop. The mechanical power input required will be the product of the force and the conductor velocity (ignoring any losses). Simultaneously, the electric power output to the load will be the product of the potential difference and the current. Ideally, if there were no power losses and no changes in stored energy in this system, the electrical power output would be equal to the mechanical power input. The machine would be acting as an ideal generator, converting mechanical energy into electrical energy.

Elementary motors. Consider now the situation where a source of electric current is attached, possibly through sliding contacts, to the loop terminals in place of the resistor so as to cause current to flow away from the observer in the left-hand side of the loop and toward the observer in the right-hand side. If the loop is rotating counterclockwise with the same velocity as before, the same potential difference will be established between the terminals (again ignoring any losses). Accordingly, there will be electrical power entering the loop equal to the product of the potential difference and the current. The current will interact with the magnetic field to produce a force, downward on the left-hand and upward on the right-hand conductor—*i.e.,* in the direction of the velocity in each case. Thus, there will be a mechanical output power equal to the force–velocity product. Again, ignoring losses and any change in stored energy, the electrical input power will be equal to the mechanical output power. The system will act as an ideal electric motor, converting electrical energy into mechanical energy.

According to the principles of mechanics, for each action there must be an equal and opposite reaction. Thus, in the system shown in Figure 44, the torque on the loop is balanced by an equal and opposite torque on the magnets and the iron yoke. The system can therefore act as a generator or a motor if the loop is held stationary and the magnet system is allowed to rotate.

Rotor and stator

Most, but not all, electric machines are based on the principles just described. A machine of this kind normally contains a rotating part, or rotor, and a stationary part, or stator. In Figure 44, the conductor loop can be fixed to the surface of a rotatable iron core to make up the rotor. In order to reduce the length of the air gaps across which the magnetic field must be produced, the conductor may in fact be imbedded in slots cut into the surface of the iron rotor. The magnetic field may be created by permanent magnets as shown or by electromagnets consisting of current-carrying coils around iron poles. In most machines, the stator is made approximately circular, with both upper and lower flux paths rather than with the one-sided yoke shown in Figure 44. The machine may equally consist of a magnet (permanent or electro-) on the rotor with conductors on the stator. For a permanent-magnet machine, this latter arrangement eliminates the need for sliding contacts or slip rings to connect the conductor loop to the external electric system.

The usual types of electric machine—induction, synchronous, and commutator—have much in common. They differ mainly in how the magnetic field is produced and how the conductors are arranged.

Other electromechanical phenomena. Other physical phenomena can be exploited to produce electromechanical energy converters, usually of a specialized nature. The force of attraction between bodies with opposite electric charges has been used in some electrostatic machines. These forces, however, are very small, even when high voltages are used. Another useful phenomenon is the

Utilization of the piezoelectric effect

piezoelectric effect in which a crystal deforms on application of an electric field. This phenomenon is utilized, for example, in an energy converter to produce underwater sound waves.

Some machines are based on the force of attraction between movable parts of an iron system that carries a magnetic field. These are commonly known as reluctance machines.

The electrical conductors in a machine need not be solid. The conductor can consist of a conducting liquid or gas. Such machines, classified as magnetohydrodynamic devices, can be used to produce electrical power (see *Magnetohydrodynamic power generators* below).

ELECTRIC GENERATORS

Electric generators, as noted above, transform mechanical power into electrical power. The mechanical power is usually obtained from a rotating shaft and is equal to the shaft torque multiplied by the rotational, or angular, velocity. The most significant generators are those used to provide power for transmission and distribution over electric power networks. The mechanical power is obtained from a number of sources: hydraulic turbines at dams or waterfalls; wind turbines; steam turbines using steam produced with heat from the combustion of fossil fuels or from the fission of heavy atomic nuclei; gas turbines burning gas directly in the turbine; or gasoline and diesel engines. The construction and the speed of the generator may vary considerably depending on the characteristics of the mechanical prime mover.

Sources of mechanical power for generators

Nearly all generators used to supply electric power networks generate alternating current, which reverses polarity at a fixed frequency (usually 50 or 60 cycles, or double reversals, per second). Since a number of generators are connected into a power network, they must operate at the same frequency for simultaneous generation. They are therefore known as synchronous generators or, in some contexts, alternators.

Synchronous generators. A major reason for selecting alternating current for power networks is that its continual variation with time allows the use of transformers. These devices convert electrical power at whatever voltage and current it is generated to high voltage and low current for long-distance transmission and then transform it down to a low voltage suitable for each individual consumer (typically 120 or 240 volts for domestic service). The particular form of alternating current used is a sine wave, which has the shape shown in Figure 45. This has been chosen because it is the only repetitive shape for which two waves displaced from each other in time can be added or subtracted and have the same shape occur as the result. The ideal is then to have all voltages and currents of sine shape. The synchronous generator is designed to produce this shape as accurately as is practical. This will become apparent as the major components and characteristics of such a generator are described below.

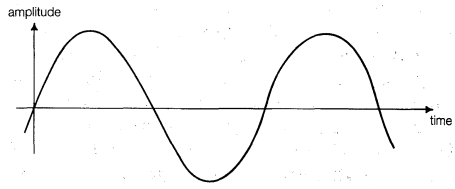


Figure 45: Sine wave.

Rotor. An elementary synchronous generator is shown in cross section in Figure 46. The central shaft of the rotor is coupled to the mechanical prime mover. The magnetic field is produced by conductors, or coils, wound into slots cut in the surface of the cylindrical iron rotor. This set of coils, connected in series, is thus known as the field winding. The position of the field coils is such that the outwardly directed or radial component of the magnetic field produced in the air gap to the stator is approximately sinusoidally distributed around the periphery of the rotor. In Figure 46, the field density in the air gap is maximum outward at the top, maximum inward at the bottom, and zero at the two sides, approximating a sinusoidal distribution.

Field winding

Stator. The stator of the elementary generator in Figure 46 consists of a cylindrical ring made of iron to provide an easy path for the magnetic flux. In this case, the stator contains only one coil, the two sides being accommodated

in slots in the iron and the ends being connected together by curved conductors around the stator periphery. The coil normally consists of a number of turns.

When the rotor is rotated, a voltage is induced in the stator coil. At any instant, the magnitude of the voltage is proportional to the rate at which the magnetic field encircled by the coil is changing with time—*i.e.*, the rate at which the magnetic field is passing the two sides of the coil. The voltage will therefore be maximum in one direction when the rotor has turned 90° from the position shown in Figure 46 and will be maximum in the opposite direction 180° later. The waveform of the voltage will be approximately of the sine form shown in Figure 45.

Frequency of electrical output and rotor speed

Frequency. The rotor structure of the generator in Figure 46 has two poles, one for magnetic flux directed outward and a corresponding one for flux directed inward. One complete sine wave is produced for each revolution of the rotor. The frequency of the electrical output, measured in hertz (cycles per second) is therefore equal to the rotor speed in revolutions per second. To provide a supply of electricity at 60 hertz, for example, the prime mover and rotor speed must be 60 revolutions per second, or 3,600 revolutions per minute. This is a convenient speed for many steam and gas turbines. For very large turbines, such a speed may be excessive for reasons of mechanical stress. In this case, the generator rotor is designed with four poles spaced at intervals of 90°. The voltage induced in a stator coil, which spans a similar angle of 90°, will consist of two complete sine waves per revolution. The required rotor speed for a frequency of 60 hertz is then 1,800 revolutions per minute. For lower speeds, a larger number of pole pairs can be used. The possible values of rotor speed, in revolutions per minute, are equal to $120f/p$, where f is the frequency and p the number of poles.

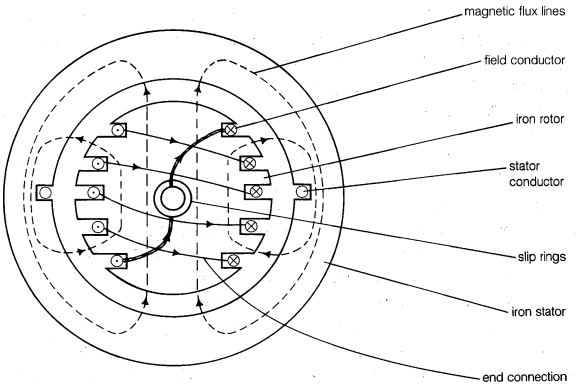


Figure 46: Elementary synchronous machine.

Stator windings. The maximum value of flux density in the air gap is limited by magnetic saturation in the stator and rotor iron, and it is typically about one tesla (weber per square metre). The effective, or root-mean-square (rms), voltage induced in one turn of a stator coil in a 60-hertz generator is about 170 volts for each metre squared of area encompassed by the coil (see below). Large synchronous generators are usually designed for a terminal voltage of several thousand volts. Each stator coil may therefore contain a number of insulated turns of conductor, and each stator winding may consist of a number of similar coils placed in sequential slots in the stator surface and connected in series as shown for the winding a-a' in Figure 47.

Phases. The voltages induced in individual coils in the distributed winding of Figure 47 are somewhat displaced in time from each other. As a result, the maximum winding voltage is somewhat less than the voltage per coil multiplied by the number of coils. The waveform is, however, still of approximately sine form. In the figure the winding a-a' spans two arcs, each of 60°. In order to make use of the whole periphery of the stator surface, two other similar windings are inserted. The voltage induced in winding b-b' will be equal in peak magnitude to that of a-a' but will be delayed in time by one-third of a cycle. The voltage in winding c-c' will be delayed by an additional third of

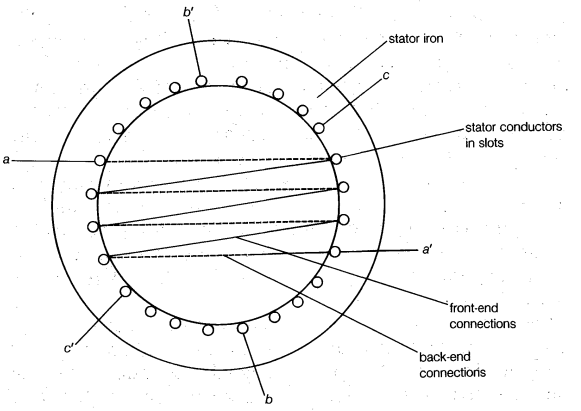


Figure 47: A three-phase winding on the stator (see text).

a cycle. This is known as a three-phase system of windings. The waveforms for the three windings, or phases, are shown in Figure 48.

Three-phase system of windings

The three-phase arrangement has a number of advantages. A single winding, or phase, requires two conductors for transmission of its electrical power to a load. At first glance, it might appear that six conductors would be required for the system in Figure 47. If, however, the waveforms of Figure 48 are considered to be those of the currents flowing in the three-phase windings, it will be seen that the sum of the three currents is zero at every instant in time. Thus, as long as the three phases are loaded equally, the terminals a', b', and c' of Figure 47 can be connected together to form a neutral point that may either be connected to ground or left open. The power of all three phases can be transmitted on three conductors. This connection is called a star, or wye, connection. Alternatively, since the three winding voltages also sum to zero at every instant, the three windings can be connected in series—a' to b, b' to c, and c' to a—to form a delta connection. The output can then be transmitted from only three conductors connected to the three junction points. Other advantages of the three-phase system will become evident in the discussion of electric motors below.

Field excitation. A source of direct current is required for the field winding, as sketched in Figure 46. In very small synchronous generators, this current may be supplied from an external source by fitting the generator shaft with two insulated copper rings, connecting the field coil ends to the rings and providing a connection to the external source through fixed carbon brushes bearing on the rings.

The power required for the field winding is that which is dissipated as heat in the winding resistance. In large generators, this is usually less than 1 percent of the generator rating, but in a generator with a capacity of 1,000 megavolt-amperes this will still be several megawatts. For most large synchronous generators, the field current is provided by another generator, known as an exciter, mounted on the same shaft. This may be a direct-current generator. In most modern installations, a synchronous generator is used as the exciter. For this purpose, the field windings of the exciter are placed on its stator and the phase windings on its rotor. A rectifier mounted on the rotating shaft is used to convert the alternating current to direct current. The field current of the main generator can then be adjusted by controlling the field current of the exciter.

Exciter

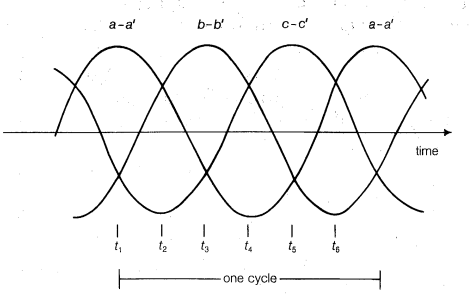


Figure 48: Waveforms of a three-phase system.

Generator rating. The capacity of a synchronous generator is equal to the product of the voltage per phase, the current per phase, and the number of phases. It is normally stated in megavolt-amperes (MVA) for large generators or kilovolt-amperes (kVA) for small generators. Both the voltage and the current are the effective, or rms, values (equal to the peak value divided by $\sqrt{2}$).

Voltage rating

The voltage rating of the generator is normally stated as the operating voltage between two of its three terminals—*i.e.,* the phase-to-phase voltage. For a winding connected in delta, this is equal to the phase-winding voltage. For a winding connected in wye, it is equal to $\sqrt{3}$ times the phase-winding voltage.

The capacity rating of the machine differs from its shaft power because of two factors—namely, the power factor and the efficiency. The power factor is the ratio of the real power delivered to the electrical load divided by the total voltage–current product for all phases. The efficiency is the ratio of the electrical power output to the mechanical power input. The difference between the two power values is the power loss consisting of losses in the magnetic iron due to the changing flux, losses in the resistance of the stator and rotor conductors, and losses from the winding and bearing friction. In large synchronous generators, these losses are generally less than 5 percent of the capacity rating. These losses must be removed from the generator by a cooling system to maintain the temperature within the limit imposed by the insulation of the windings.

High-speed synchronous generators. Generators driven by high-speed turbines are almost always constructed with horizontal shafts. The rotor diameter is usually limited to a maximum of about one metre because of the high centrifugal forces produced. The length of the rotor may be several metres. The rotor shaft and the field structure are made of a solid alloy steel forging in which slots are machined to accept the field coils, as shown in Figure 46. These coils are insulated typically with mica and glass laminate. The coils are held in place by nonmagnetic wedges in the tops of the slots.

The stator provides a path for the continuously varying magnetic flux. The stator core is therefore constructed of thin sheets, or laminations, of magnetic steel. The steel, being an electrical conductor, would tend to short-circuit the voltage induced in it if it were solid. Lamination breaks up the path along the length of the stator and keeps the power losses in the stator steel at an acceptable value. Slots are punched around the inside periphery of the laminations to accommodate the stator coils. In large generators, each stator coil normally contains only one turn.

High-speed generators are enclosed within a closed cylindrical stator housing that extends between the bearings at the two ends. They are cooled by hydrogen gas circulating within the housing and also frequently through ducts within the stator conductors. Very large generators are cooled by circulating water through the stator and rotor conductors.

The ratings of synchronous generators for large power systems extend up to about 2,000 megavolt-amperes. Smaller power systems use generators of lower rating (*e.g.,* 50 megavolt-amperes and up) since it is usually not desirable to have more than 10 percent of the total required system generation in one machine.

Waterwheel generators. Hydraulic turbines are of various types, the choice depending largely on the height of water fall and on the power rating (see *Water turbines* above). The range of speed for which hydraulic turbines give acceptable efficiency is much lower than for steam turbines. The rotational speed is generally in the range of 60 to 720 revolutions per minute. The construction of low-speed synchronous generators is substantially different from that of high-speed units. To produce power at 60 hertz, the number of rotor poles is in the range of 10 to 120 for the above speed range. For these machines the rotor poles are of the projecting, or salient, type. Figure 49 shows two poles of a 12-pole generator. Each pole, made of laminated magnetic steel, is encircled by a field coil. The pole is shaped so as to make the air-gap magnetic field distribution approximately sinusoidal.

Large hydraulic generators may have individual ratings in excess of 200 megavolt-amperes. They are mounted with a vertical shaft directly coupled to the turbine. The combination is usually supported on a single bearing, either above or below. The diameter is made relatively large to obtain a high peripheral velocity at low rotational speeds. The axial length of the generator is relatively short. The windings are frequently water-cooled. The rotor has to be designed to withstand a considerable overspeed condition that may arise if the generator loses its electrical load and there is a significant time delay in cutting off the water flow to the turbine.

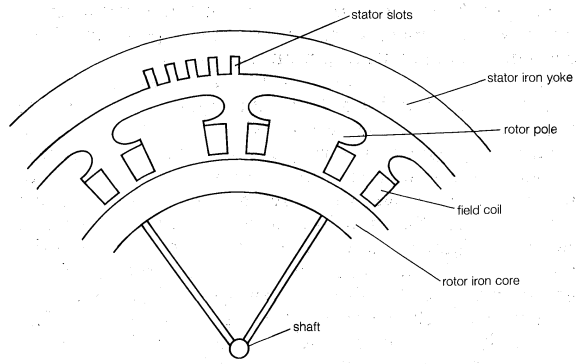


Figure 49: A two-pole cross section of a 12-pole, low-speed synchronous generator.

Generators for motor vehicles. Such vehicles as automobiles, buses, and trucks require a direct-voltage supply for ignition, lights, fans, and so forth. In modern vehicles the electric power is generated by an alternator mechanically coupled to the engine. The alternator normally has a rotor field coil supplied with current through slip rings. The stator is fitted with a three-phase winding. A rectifier is used to convert the power from alternating to direct form. A regulator is used to control the field current so that the output voltage of the alternator-rectifier is properly matched to the battery voltage as the speed of the engine varies.

Alternator-rectifier

Permanent-magnet generators. In small ratings, the magnetic field of the synchronous generator may be provided by permanent magnets. The rotor structure can consist of a ring of magnetic iron with magnets mounted on its surface, as in the four-pole structure shown in Figure 50. A magnet material such as neodymium-boron-iron or samarium-cobalt can provide a magnetic flux density in the air gap comparable to that produced with field windings, using a radial depth of magnet of about 10 millimetres. Other magnet materials such as ferrite can be used, but with a considerable reduction in air-gap flux density and a corresponding increase in generator dimensions.

Permanent-magnet generators are simple in that they require no system for the provision of field current. They are highly reliable. They do not, however, contain any means for controlling the output voltage, and this may vary with changes in load.

Induction generators. An induction machine (see *Induction motors* below) can operate as a generator if it is connected to an electric supply network operating at a substantially constant voltage and frequency. If torque is applied to the induction machine by a prime mover, it will tend to rotate somewhat faster than its synchronous

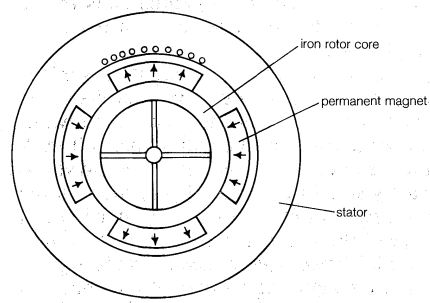


Figure 50: Cross section of a permanent-magnet generator.

speed, which is equal to 120 f/p revolutions per minute, where f is the supply frequency and p is the number of poles in the machine. The rotor conductors, moving faster than the air-gap field, will have induced currents that interact with the field to produce a torque with which to balance that applied by the prime mover. A stator current will then flow into the supply network delivering electrical power. The amount of power delivered is approximately proportional to the difference between the rotor speed and the field speed. This difference is typically of the order of 0.5 to 2 percent of rated speed at rated load.

An induction generator cannot normally provide an independent electrical power source because it does not contain a source of its own magnetic field. Stand-alone induction generators can, however, operate with the aid of appropriate loading capacitors.

Induction generators are frequently preferred over synchronous generators for small and remote hydroelectric sites because they are not subject to loss of synchronism following transient changes in the power system.

Special type of synchronous generator

Inductor alternators. An inductor alternator is a special kind of synchronous generator in which both the field and the output winding are on the stator. In the homopolar type of machine, the magnetic flux is produced by direct current in a field coil concentric with the shaft. In the heteropolar type, the field coils are in slots in the stator.

Voltage is generated in the output windings by pulsations in the flux in individual stator teeth. These pulsations are produced by use of a toothed rotor, which causes the reluctance of the air path from the rotor to each stator tooth to vary periodically with rotation.

Inductor alternators are useful as high-frequency generators. They also are useful in situations requiring high reliability, a feature achieved by their having no electrical connections to the rotor.

Direct-current generators. A direct-current (DC) generator is a rotating machine that supplies an electrical output with unidirectional voltage and current. The basic principles of operation are the same as those for synchronous generators. Voltage is induced in coils by the rate of change of the magnetic field through the coils as the machine rotates. This induced voltage is inherently alternating in form since the coil flux increases and then decreases, usually with a zero average value.

The field is produced by direct current in field coils or by permanent magnets on the stator. The output, or armature, windings are placed in slots in the cylindrical iron rotor. A simplified machine with only one rotor coil is shown in Figure 51. The rotor is fitted with a

Commutator

mechanical rotating switch, or commutator, that connects the rotor coil to the stationary output terminals. This commutator reverses the connections at the two instants in each rotation when the rate of change of flux in the coil is zero—*i.e.*, when the enclosed flux is maximum (positive) or minimum (negative). The output voltage is then unidirectional but is pulsating for the single case of one rotor coil. In practical machines, the rotor contains many coils symmetrically arranged in slots around the periphery and all connected in series. Each coil is connected to a segment on a multi-bar commutator. In this way, the output voltage consists of the sum of the induced voltages in a number of individual coils displaced around half the periphery. The magnitude of the output voltage is then approximately constant, containing only a small ripple due to the limited number of coils. The voltage magnitude is proportional to the rotor speed and the magnetic flux. Control of output voltage is normally provided by control of the direct current in the field.

For convenience in design, direct-current generators are usually constructed with four to eight field poles, partly to shorten the end connections on the rotor coils and partly to reduce the amount of magnetic iron needed in the stator. The number of stationary brushes bearing on the rotating commutator is usually equal to the number of poles but may be only two in some designs.

The field current for the generator may be obtained from an external source, such as a battery or a rectifier, as shown in Figure 52A. In this case, the generator is classed as separately excited. Alternatively, it may be noted that

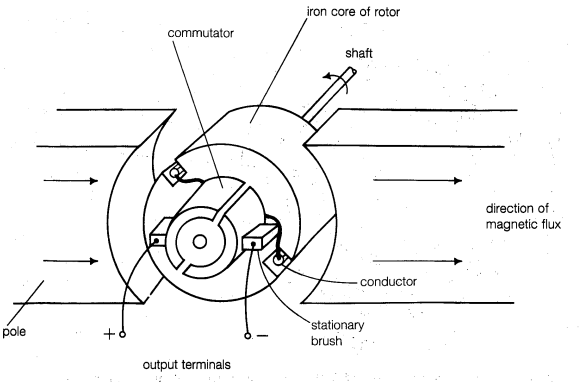


Figure 51: Direct-current generator.

the output of the DC generator is unidirectional and therefore may be used as a source to supply its own field current, as shown in Figure 52B. In this case, the generator is referred to as shunt-excited. It has the advantage of requiring no independent supply. Residual magnetic flux in the iron poles produces a small generated voltage when the machine is brought up to speed. This causes a field current that increases the flux and in turn the generated voltage. The voltage builds up until saturation in the iron limits the voltage produced. The stable value of generated voltage can be adjusted over a limited range by adjusting the value of a resistor placed in series with the field coil across the output terminals.

Shunt-excited DC generator

Direct-current generators were widely used prior to the availability of economical rectifier systems supplied by alternators. For example, they were commonly employed for charging batteries and for electrolytic purposes. In some applications, the direct-current generator retains an advantage over the alternator-rectifier in that it can operate as a motor as well, reversing the direction of power flow. An alternator, by contrast, must be fitted with a more complex rectifier-inverter system to accomplish power reversal.

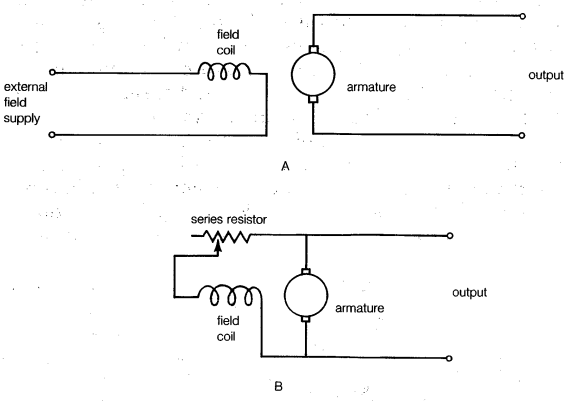


Figure 52: *Types of direct-current generators on the basis of source of field current.*
(A) Separately excited DC generator and (B) shunt-excited DC generator (see text).

ELECTRIC MOTORS

Electric motors transform electrical power into mechanical power. In most instances, the electrical power is obtained from a power distribution network through appropriate control apparatus. In special situations, the electric supply may come from a battery, as, for example, in an automobile.

The basic principles of electric motors were discussed above. Most motors develop their mechanical torque by the interaction of conductors carrying current in a direction at right angles to a magnetic field. The various types of electric motor differ in the ways in which the conductors and the field are arranged and also in the control that can be exercised over mechanical output torque, speed, and position. Each of the major kinds is delineated below.

Differences between motor types

Induction motors. The simplest type of induction motor is shown in cross section in Figure 53. A three-phase

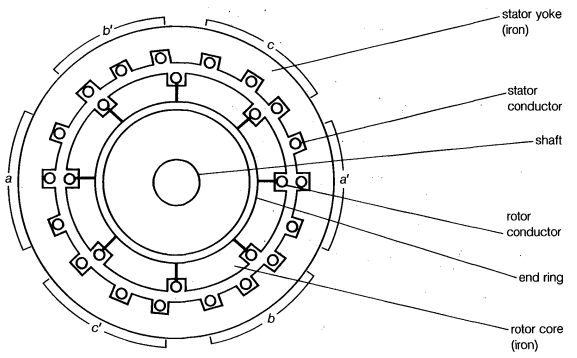


Figure 53: Cross section of a three-phase induction motor.

set of stator windings is inserted in slots in the stator iron. These windings may be connected either in a wye configuration, normally without external connection to the neutral point, or in a delta configuration. The rotor consists of a cylindrical iron core with conductors placed in slots around the surface. In the most usual form, these rotor conductors are connected together at each end of the rotor by a conducting end ring.

The basis of operation of the induction motor may be developed by first assuming that the stator windings are connected to a three-phase electric supply and that a set of three sinusoidal currents of the form shown in Figure 48 flow in the stator windings. Figure 54 shows the effect of the currents in producing a magnetic field across the air gap of the machine. For simplicity, only the central conductor loop for each phase winding is shown. At the instant t_1 in Figure 48, the current in phase a is maximum positive, while that in phases b and c is half that value negative. The result is a magnetic field with an approximately sinusoidal distribution around the air gap with a maximum outward value at the top and a maximum inward value at the bottom. At time t_2 in Figure 48 (*i.e.*, one-sixth of a cycle later), the current in phase c is maximum negative, while that in both phase b and phase a is half value positive. The result, as shown for t_2 in Figure 54 is again a sinusoidally distributed magnetic field but rotated 60° counterclockwise. Examination of the current distribution for t_3, t_4, t_5, and t_6 shows that the magnetic field continues to rotate as time progresses. The field completes one revolution in one cycle of the stator current. Thus, the combined effect of three sinusoidal currents, uniformly displaced in time and flowing in three stator windings uniformly displaced in angular position, is to produce a rotating magnetic field with a constant magnitude and a mechanical angular velocity that depends on the frequency of the electric supply.

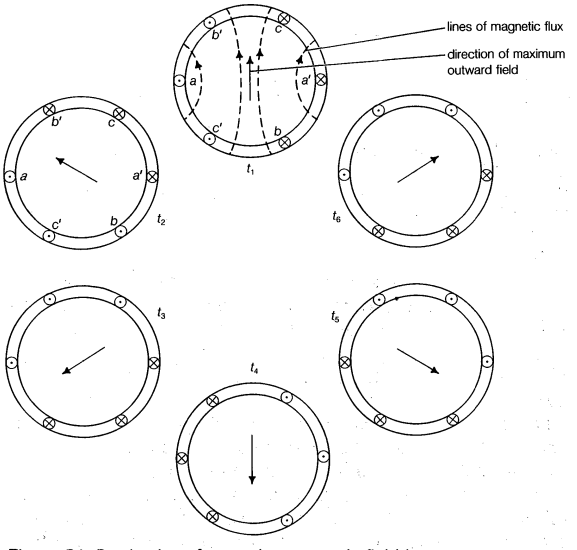


Figure 54: *Production of a rotating magnetic field by three-phase currents in three stator windings.* The symbol ⊙ indicates current flow toward observer; ⊗ denotes current flow away. The t represents the time instants in Figure 48.

The effect of the rotating field on the rotor

The rotational motion of the magnetic field with respect to the rotor conductors causes a voltage to be induced in each, proportional to the magnitude and the velocity of the field relative to the conductors. Since the rotor conductors are short-circuited together at each end, the effect will be to cause currents to flow in these conductors. In the simplest mode of operation, these currents will be about equal to the induced voltage divided by the conductor resistance. The pattern of rotor currents for the instant t_1 of Figure 54 is shown in Figure 55. The currents are seen to be sinusoidally distributed around the rotor periphery and to be located so as to produce a counterclockwise torque on the rotor (*i.e.*, a torque in the same direction as the field rotation). This torque acts to accelerate the rotor and to rotate the mechanical load. As the rotational speed of the rotor increases, its speed relative to that of the rotating field decreases. Thus, the induced voltage is reduced, leading to a proportional reduction in rotor conductor current and in torque. The rotor speed reaches a steady value when the torque produced by the rotor currents equals the torque required at that speed by the load with no excess torque available for accelerating the combined inertia of the load and the motor.

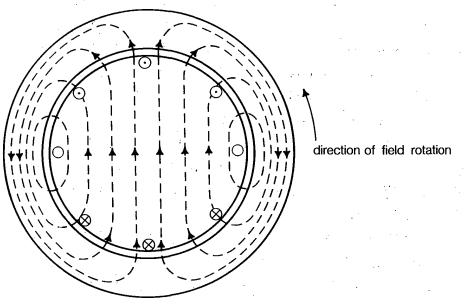


Figure 55: A rotating field and the currents that it produces in shorted rotor conductors.

The mechanical output power must be provided by an electrical input power. The original stator currents shown in Figure 54 are just sufficient to produce the rotating magnetic field. To maintain this rotating field in the presence of the rotor currents of Figure 55, it is necessary that the stator windings carry an additional component of sinusoidal current of such a magnitude and phase as to cancel the effect of the magnetic field that would otherwise be produced by the rotor currents in Figure 55. The total stator current in each phase winding is then the sum of a sinusoidal component to produce the magnetic field and another sinusoid, leading the first by one-quarter of a cycle, or 90°, to provide the required electrical power. The second, or power, component of the current is in phase with the voltage applied to the stator, while the first, or magnetizing, component lags the applied voltage by a quarter cycle, or 90°. At rated load, this magnetizing component is usually in the range of 0.4 to 0.6 of the magnitude of power component.

Power and magnetizing components of current

A majority of three-phase induction motors operate with their stator windings connected directly to a three-phase electric supply of constant voltage and constant frequency. Typical supply voltages range from 230 volts line-to-line for motors of relatively low power (*e.g.*, 0.5 to 50 kilowatts) to about 15 kilovolts line-to-line for high-power motors up to about 10 megawatts.

Except for a small voltage drop in the resistance of the stator winding, the supply voltage is matched by the time rate of change of the magnetic flux in the stator of the machine. Thus, with a constant-frequency, constant-voltage supply, the magnitude of the rotating magnetic field is held constant, and the torque is roughly proportional to the power component of the supply current.

With the induction motor shown in Figures 53, 54, and 55, the magnetic field rotates through one revolution for each cycle of the supply frequency. For a 60-hertz supply, the field speed is then 60 revolutions per second, or 3,600 per minute. The rotor speed is less than the speed of

the field by an amount that is just enough to induce the required voltage in the rotor conductors to produce the rotor current needed for the load torque. At full load, the speed is typically 0.5 to 5 percent lower than the field speed (often called synchronous speed), with the higher percentage applying to smaller motors. This difference in speed is frequently referred to as the slip.

Other synchronous speeds can be obtained with a constant frequency supply by building a machine with a larger number of pairs of magnetic poles, as opposed to the two-pole construction of Figure 53. The possible values of magnetic-field speed in revolutions per minute are $120f/p$, where f is the frequency in cycles per second and p is the number of poles (which must be an even number). A given iron frame can be wound for any one of several possible numbers of pole pairs by using coils that span an angle of approximately $(360/p)°$. The torque available from the machine frame will remain unchanged, since it is proportional to the product of the magnetic field and the allowable coil current. Thus, the power rating for the frame, being the product of torque and speed, will be roughly inversely proportional to the number of pole pairs. The most common synchronous speeds for 60-hertz motors are 1,800 and 1,200 revolutions per minute.

Construction of induction motors. The stator frame consists of laminations of silicon steel, usually with a thickness of about 0.5 millimetre. Lamination is necessary since a voltage is induced along the axial length of the steel as well as in the stator conductors. The laminations are insulated from each other by a varnish layer in most cases. This breaks up the conducting path in the steel and limits the losses (known as eddy current losses) in the steel.

The stator coils are normally made of copper; round conductors of many turns per coil are used for small motors, and rectangular bars of fewer turns are employed for larger machines. The coils are electrically insulated. It is common practice to bring only three leads out to a terminal block whether the winding is connected in wye or in delta.

The magnetic part of the rotor is also made of steel laminations, mainly to facilitate stamping conductor slots of the desired shape and size. In most induction motors, the rotor winding is of the squirrel-cage type where solid conductors in the slots are shorted together at each end of the rotor iron by conducting end rings. In such machines there is no need to insulate the conductors from the iron. For motors up to about 300 kilowatts, the squirrel cage often consists of an aluminum casting incorporating the conductors, the end rings, and a cooling fan. For larger motors, the squirrel cage is made of copper, aluminum, or brass bars welded or brazed to end rings of a similar material. In any case, the rotor is very rugged and is also economical to produce in contrast to rotors requiring an electrically insulated winding.

Squirrel-cage rotor winding

The rotor slots need not be rectangular. The shape of the slots can be designed to provide a variety of torque-speed characteristics.

Starting characteristics. When operated from a constant-frequency supply, the three-phase induction motor constitutes essentially a constant-speed drive, with the speed decreasing only 1 to 5 percent as load torque is increased from zero to rated value. In most installations, induction motors can be started and brought up to speed by connecting the stator terminals directly to the electric supply. This establishes the rotating field in the machine. At zero speed the velocity of this field, relative to that of the rotor, is high. If the rotor current were limited only by the resistance of the rotor bars, the rotor currents would be extremely high. The starting current is, however, limited by additional paths for the magnetic field around the stator and rotor conductors, known as flux leakage paths. Usually, the starting current is thus limited to about four to seven times rated current when started on full voltage. The torque at starting is usually in the range of 1.75 to 2.5 times rated value.

If the stator current on starting is larger than is permissible from the electric supply system, the motor may be started on a reduced voltage of about 70 to 80 percent using a step-down transformer. Alternatively, the stator windings can be connected in wye to start and can be switched to delta as the speed approaches rated value. Such measures reduce the starting torque substantially. A reduction in the starting voltage to 75 percent results in a reduction in the electric supply current to 56 percent but also results in only 56 percent of the starting torque that would be provided with full voltage.

Other motor starters insert a resistance or inductance in series with each stator phase during the starting period.

Protection. The heat generated by power losses in the conductors and iron parts of the machine, as well as the friction heat, must be removed by the cooling system to limit the temperature of the motor. The main purpose of protection apparatus is to prevent damage to the most vulnerable part of the motor, the insulation on the windings. For low-power motors, a temperature-sensitive device is often mounted inside the motor and used to switch off the electric supply if the temperature reaches its limiting value. With larger motors, temperature-sensitive detectors may be imbedded at one or more locations in the stator windings.

Wound-rotor induction motors. Some special induction motors are constructed with insulated coils in the rotor similar to those in the stator winding. The rotor windings are usually of a three-phase type with three connections made to insulated conducting rings (known as slip rings) mounted on an internal part of the rotor shaft. Carbon brushes provide for external electric connections.

Use of slip rings

A wound-rotor motor with three resistors connected to its slip rings can provide a high starting torque without excessive starting current. By varying the resistance, a degree of speed control can be provided for some types of mechanical load. The efficiency of such drives is, however, low unless the speed is reasonably close to the synchronous value because of the high losses in the rotor circuit resistances. As an alternative, an electronic rectifier-inverter system can be connected to the rotor slip rings to extract power and feed it back to the electric supply system. This arrangement, normally called a slip recovery system, provides speed control with acceptable efficiency.

Single-phase induction motors. The development of a rotating field in an induction machine requires a set of currents displaced in phase (as shown in Figure 48) flowing in a set of stator windings that are displaced around the stator periphery. While this is straightforward where a three-phase supply is available, most commercial and domestic supplies are only of a single phase, typically with a voltage of 120 or 240 volts. There are several ways in which the necessary revolving field can be produced from this single-phase supply.

Capacitor induction motor. This motor is similar to the three-phase motor except that it has only two windings on its stator displaced 90° from each other. One winding (a-a' in Figure 56) is connected directly to the single-phase supply. For starting, the other winding (commonly called the auxiliary winding) is connected through a capacitor (a device that stores electric charge) to the same supply. The effect of the capacitor is to make the current entering the winding b-b' lead the current in a-a' by approximately 90°, or one-quarter of a cycle, with the rotor at standstill. Thus, the rotating field and the starting torque are provided.

Means of providing a rotating field

As the motor speed approaches its rated value, it is no longer necessary to excite the auxiliary winding to maintain the rotating field. The currents produced in the rotor

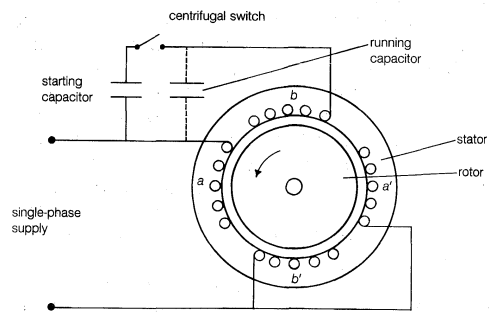


Figure 56: Capacitor induction motor.

squirrel-cage bars as they pass the winding *a-a′* are retained with negligible change as they rotate past the winding *b-b′*. The rotor can continue to generate the rotating field with only winding *a-a′* connected. The winding *b-b′* is usually disconnected by a centrifugal switch that opens when the speed is about 80 percent of rated value.

Power ratings for these capacitor-start induction motors are usually restricted to about two kilowatts for a 120-volt supply and 10 kilowatts for a 230-volt supply because of the limitations on the voltage drop in the supply lines, which would otherwise occur on starting. Typical values of synchronous speed on a 60-hertz supply are 1,800 or 1,200 revolutions per minute for four- and six-pole motors, respectively. Lower-speed motors can be constructed with more poles but are less common.

The efficiency of the motor can be somewhat increased and the line current decreased by the use of two capacitors, only one of which is taken out of the circuit (by means of a centrifugal switch) as the rated speed is approached. The remaining capacitor continues to provide a leading current to phase *b-b′*, approximating a two-phase supply. This arrangement, also shown in Figure 56, is known as a capacitor-start, capacitor-run motor.

Capacitor induction motors are widely used for heavy-duty applications requiring high starting torque. Examples are refrigerator compressors, pumps, and conveyors.

Use of two stator windings

Split-phase motors. An alternative means of providing a rotating field for starting is to use two stator windings, as in Figure 57, where the auxiliary winding *b-b′* is made of more turns of smaller conductors so that its resistance is much larger than that of winding *a-a′*. The effect of this is that the current in phase *b-b′* leads that of *a-a′*, but only by about 20–30 degrees at standstill. While the field is largely pulsating, it contains enough rotating component to provide a starting torque of 1.5 to 2.0 times rated value. To prevent overheating, the auxiliary winding is disconnected by a centrifugal switch when the speed reaches 75–80 percent of rated value.

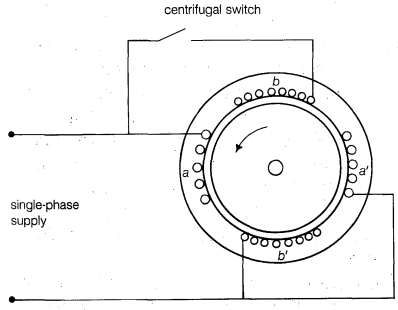


Figure 57: Split-phase induction motor.

These split-phase motors are inexpensive to produce and are installed in many domestic appliances. Where more than one steady speed is required, as in household laundry appliances, the motor may be wound for two alternative pole pairs, one for low speed and the other for high speed.

Shaded-pole motors. The shaded-pole motor is provided with a main winding connected to the single-phase electric supply. In addition, it has a permanently short-circuited winding located ahead of the main winding in the direction of rotation. This second winding is known as a shading coil and consists of one or more shorted turns. The shading coil delays the establishment of magnetic flux in the region that it encircles and thus produces a small component of rotating field at standstill.

The starting torque is small, typically only 30 to 50 percent of the rated torque. As a result, the motor is suitable only for mechanical loads, such as fans, for which the torque is low at low speed and increases with speed.

Shaded-pole motors are inefficient because of the losses in the permanently shorted winding. As a result, they are used only in small power ratings where efficiency is less important than initial cost. Typical efficiencies are up to 30 percent in larger units and less than 5 percent in very small ones. They are used mainly for fans and other small household appliances.

Servomotors. A servomotor is a small induction motor with two stator windings displaced 90° with respect to each other around its periphery. The rotor is usually of the squirrel-cage type but made with relatively high resistance conductors. The purpose of the motor is to provide a controlled torque in either direction of operation. To achieve this, one winding is connected to a single-phase, constant-frequency supply. The other winding is provided with a voltage of the same frequency, displaced 90° in phase. This voltage is normally provided by an electronic amplifier with a low power signal input. The motor torque is approximately proportional to the voltage on this second winding and thus to the signal input. The direction of the torque can be reversed by changing the input signal from 90° leading to 90° lagging.

On some servomotors the rotor consists of an aluminum cup fitted in the air gap between the stator and a stationary iron core. This rotor has low inertia and is capable of high acceleration. Servomotors are made only in small power ratings because of their high losses and low efficiency. They are used in position-control systems.

Linear induction motors. A linear induction motor provides linear force and motion rather than rotational torque. The shape and operation of a linear induction motor can be visualized as depicted in Figure 58 by making a radial cut in a rotating induction machine and flattening it out. The result is a flat "stator," or upper section, of iron laminations that carry a three-phase, multipole winding with conductors perpendicular to the direction of motion. The "rotor," or lower section, could consist of iron laminations and a squirrel-cage winding but more normally consists of a continuous copper or aluminum sheet placed over a solid or laminated iron backing.

Use of linear motors in rapid-transit vehicles

An emerging application of linear motors is in rapid-transit vehicles for public transportation. The stator (as described above) is carried on the underside of the vehicle, and the rotor is located between the rails on the track. An advantage of this type of propulsion is that high acceleration and braking can be obtained without dependence on adhesion of the steel wheels to the steel rails in the presence of rain, ice, or a steep slope.

Electrical power is supplied to such a rapid-transit vehicle through sliding connections to an energized rail or overhead wire. To provide speed control and braking, an electronic power-conditioning apparatus on board the vehicle produces a three-phase output of the desired voltage and frequency.

In an alternative arrangement for vehicle propulsion, the copper and iron sheets of Figure 58 can be placed on the underside of the vehicle and sections of stator can be placed at intervals along the track. This has the advantage that no electric power need be supplied to the vehicle itself.

Linear induction motors also are used to drive conveyors, sliding doors, textile shuttles, and machine tools. Their advantage is that no physical contact is required and thus wear and maintenance are minimized. In another form, linear motors are used as electromagnetic pumps where the rotor consists of a conducting fluid, such as a liquid metal (say, mercury of sodium-potassium alloy).

The efficiency of linear motors is somewhat less than

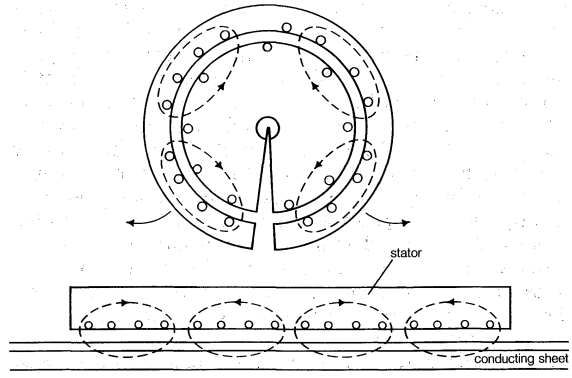


Figure 58: *Evolution of a linear induction motor.*
The four-pole induction motor is shown as (top) split open and (bottom) flattened (see text).

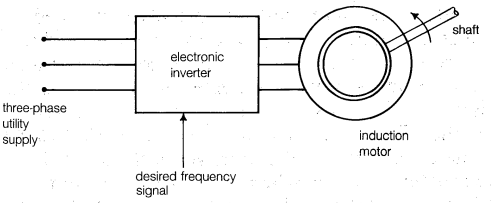


Figure 59: A variable-frequency, variable-speed induction-motor drive system.

that of rotating motors because of end effects. Its "rotor" must be magnetized as it comes under the "stator." This reduces the effectiveness of the first one or two pole spans. The input current is also relatively high because the air gap is usually larger than in rotating machines and more current is required to produce the magnetic field across it.

Induction motors for speed and position control. On a constant-frequency supply, an induction motor is essentially a near-constant speed drive. Induction motors, however, can be used to provide accurate speed and position control in either direction of rotation by furnishing a controllable-voltage, controllable-frequency three-phase supply. This is done by means of an electronic inverter, as shown in Figure 59. Using semiconductor switches (*e.g.*, transistors or thyristors), the utility supply is converted into a set of three near-sinusoidal inputs of controlled voltage and frequency to the stator winding. The speed of the motor will then approach the synchronous value of $120\,f/p$ revolutions per minute for a controlled frequency of f cycles per second. Reversal of the phase sequence from *abc* to *acb* reverses the direction of the torque. For accurate control of speed or of position, the speed of the shaft can be monitored by a tachometer or position sensor and compared with a signal representing the desired value. The difference is then used to control the inverter frequency. Generally, the voltage varies directly with the frequency to keep the magnitude of the magnetic field constant.

Synchronous motors. Such a motor is one in which the rotor normally rotates at the same speed as the revolving field in the machine. The stator is similar to that of an induction machine (as in Figure 53) consisting of a cylindrical iron frame with windings, usually three-phase, located in slots around the inner periphery. The difference is in the rotor, which normally contains an insulated winding connected through slip rings or other means to a source of direct current (see Figure 46).

The principle of operation of a synchronous motor can be understood by considering the stator windings to be connected to a three-phase alternating-current supply. The effect of the stator current is to establish a magnetic field rotating at $120\,f/p$ revolutions per minute for a frequency of f hertz and for p poles. A direct current in a p-pole field winding on the rotor will also produce a magnetic field rotating at rotor speed. These two magnetic fields will tend to align with each other. With no load torque, they may be assumed to be in alignment. As mechanical load is applied, the rotor slips back a number of degrees with respect to the rotating field of the stator, developing torque and continuing to be drawn around by this rotating field. The angle between the fields increases as load torque is increased. The maximum available torque is achieved for given magnitudes of stator and rotor currents when the angle by which the rotor field lags the stator field is 90°. Application of more load torque will stall the motor.

One advantage of the synchronous motor is that the magnetic field of the machine can be produced by the direct current in the field winding, so that the stator windings need to provide only a power component of current in phase with the applied stator voltage—*i.e.*, the motor can operate at unity power factor. This condition minimizes the losses and heating in the stator windings.

The power factor of the stator electrical input can be directly controlled by adjustment of the field current. If the field current is increased beyond the value required to provide the magnetic field, the stator current changes to include a component to compensate for this overmagnetization. The result will be a total stator current that leads the stator voltage in phase, thus providing to the power system reactive volt-amperes needed to magnetize other apparatuses, such as transformers and induction motors. Operation of a large synchronous motor at such a leading power factor may be an effective way of improving the overall power factor of the electrical loads in a manufacturing plant to avoid additional electric supply rates that may otherwise be charged for low power-factor loads.

Applications of three-phase synchronous motors

Three-phase synchronous motors find their major application in industrial situations where there is a large, reasonably steady mechanical load, usually in excess of 300 kilowatts, and where the ability to operate at leading power factor is of value. Below this power level, synchronous machines are generally more expensive than induction machines. In some instances, a synchronous machine is installed for the sole purpose of improving overall plant power factor. In this case, it is called a synchronous capacitor because it provides the same power factor correction as capacitors connected across the supply line.

The field current may be supplied from an externally controlled rectifier through slip rings, or, in larger motors, it may be provided by a shaft-mounted rectifier with a rotating transformer or generator.

A synchronous motor with only a field winding carrying a direct current would not be self-starting. At any speed other than synchronous speed, its rotor would experience an oscillating torque of zero average value as the rotating magnetic field repeatedly passes the slower moving rotor. Normally, a short-circuited winding similar to that of an induction machine is added to the rotor to provide starting torque, as shown in Figure 60. The motor is started, either with full or reduced stator voltage, and brought up to about 95 percent of synchronous speed, usually with the field winding short-circuited to protect it from excessive induced voltage. The field current is then applied and the rotor pulls into synchronism with the revolving field.

This additional rotor winding is usually referred to as a damper winding because of its additional property of damping out any oscillation that might be caused by sudden changes in the load on the rotor when in synchronism. Adjustment to load changes involves changes in the angle by which the rotor field lags the stator field and thus involves short-term changes in instantaneous speed. These cause currents to be induced in the damper windings, producing a torque that acts to oppose the speed change.

Protection for synchronous motors is similar to that employed with large induction motors. Temperature may be sensed in both the stator and field windings and used to switch off the electric supply. Considerable heating occurs in the rotor-damper winding during starting, and a timer is frequently installed to prevent repeated starts within a limited time interval.

Permanent-magnet motors. The magnetic field for a synchronous machine may be provided by using permanent magnets rather than a field winding. This eliminates the need for slip rings and an external source of field current and provides a simple rugged rotor. The motor does not, however, have a means of controlling the stator power factor.

Use in variable-speed drives

The rotor can be of the form shown in Figure 50 with radially directed magnets made of neodymium-boron-iron, samarium-cobalt, or ferrite. The machine in the figure does not contain a damper winding and thus cannot be started on a constant-frequency supply. The main application for a motor of this type is in variable-speed drives

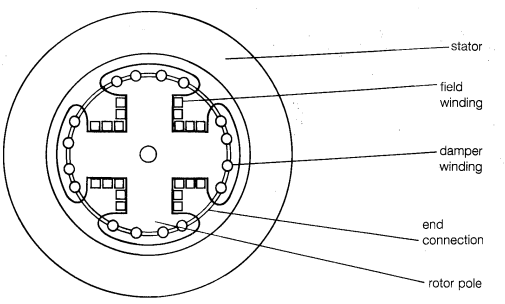


Figure 60: Cross section of rotor of a four-pole synchronous motor.

where the stator is supplied from a variable-frequency, variable-voltage source. Where starting capability is required, the magnets are imbedded in the rotor iron and a damper winding is placed in slots in the rotor surface (see Figure 60).

An alternative form of permanent-magnet motor is shown in Figure 61. Circumferentially directed magnets provide flux to iron poles, which in turn set up a radial field in the air gap. This form is particularly suitable for small motors using ferrite magnets.

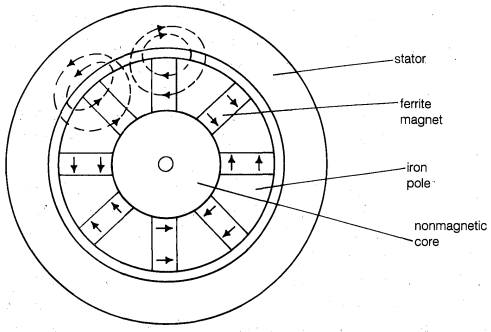


Figure 61: Cross section of an eight-pole synchronous motor with circumferentially directed permanent magnets.

Hysteresis motors. A distinctive feature of synchronous motors is that the speed is uniquely related to the supply frequency. As a result, several special types of synchronous motors have found wide application in devices such as clocks, tape recorders, and phonographs. One of the most extensively used is the hysteresis motor in which the rotor consists of a ring of a semi-permanent magnet material like a high-carbon steel. At full speed, the motor operates as a permanent-magnet machine. If the speed is reduced by pulling the rotor out of synchronism, the stator field causes the rotor material to be cyclically magnetized around its hysteresis loop, resulting in a rotor field that lags the stator field by a few degrees and continues to produce torque. These motors provide good starting torque and are very quiet. Their efficiency is low, and applications are restricted to small power ratings.

Reluctance motors. Machines of this kind operate on the principle that forces are established tending to minimize the volume of any air gap in an iron system carrying a magnetic field. One of the forms of a reluctance motor is shown in cross section in Figure 62. The rotor consists of four iron poles with no electrical windings. The stator has six poles each with a current-carrying coil. In the condition represented in the figure, current has just been passed through coils a and a', producing a torque on the rotor aligning two of its poles with those of the a-a' stator. The current is now switched off in coils a and a' and switched on to coils b and b'. This produces a counterclockwise torque on the rotor aligning two rotor poles with stator poles b and b'. This process is then repeated with stator coils c and c' and then with coils a and a'. The torque is dependent on the magnitude of the coil currents but is independent of its polarity. The direction of rotation can be changed by changing the order in which the coils are energized.

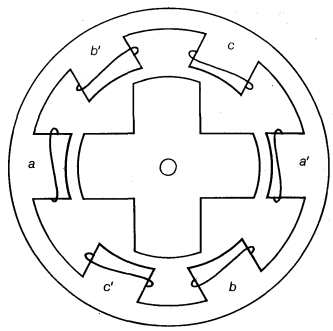


Figure 62: A reluctance motor in cross section.

The currents in the stator coils are usually controlled by semiconductor switches connecting the coils to a direct voltage source. A signal from a position sensor mounted on the motor shaft is used to activate the switches at the appropriate time instants. Frequently a magnetic sensor based on the Hall effect is employed. (The Hall effect involves the development of a transverse electric field in a semiconductor material when it carries a current and is placed in a magnetic field perpendicular to the current.) The overall system is known as a self-synchronous motor drive. It can operate over a wide and controlled speed range.

Self-synchronous motor drive

There are several other configurations for reluctance motors. In one form, the rotor consists of an iron ring with radial cuts or slots through it. A p-pole rotor has p sectors, or arcs. The magnetic flux travels circumferentially around the arc of this rotor ring, completing the path between adjacent stator poles.

In another form, the rotor has salient poles of the configuration shown in Figure 60 but without the field windings. The stator is cylindrical and contains a three-phase winding connected to a constant-frequency supply. A damper winding is fitted in the rotor surface so that the machine can start as an induction motor. After the rotor pulls into synchronism with the rotating field of the stator, it operates as a synchronous motor at constant speed.

Single-phase synchronous motors. A revolving field can be produced in synchronous motors from a single-phase source by use of the same method as for single-phase induction motors. With the main stator winding connected directly to the supply, an auxiliary winding may be connected through a capacitor, as in Figure 56. Alternatively, an auxiliary winding of a higher resistance can be employed, as in Figure 57. For small clock motors, the shaded-pole construction of the stator is widely used in combination with a hysteresis-type rotor (see above). The efficiency of these motors is very low, usually less than 2 percent, but the cost is low as well.

Direct-current commutator motors. A sketch of an elementary form of DC motor is provided in Figure 51. A stationary magnetic field is produced across the rotor by poles on the stator. These poles may be encircled by field coils carrying direct current, or they may contain permanent magnets. The rotor or armature consists of an iron core with a coil accommodated in slots. The ends of the coil are connected to the bars of a commutator switch mounted on the rotor shaft. Stationary graphite brushes lead to external terminals.

Components and characteristics

Suppose a direct-current supply is connected to the armature terminals such that a current enters at the positive terminal shown in Figure 51. This current interacts with the magnetic flux to produce a counterclockwise torque, which in turn accelerates the rotor. When the rotor has turned about 120° from the position shown in the figure, the connection from the supply to the armature coil is reversed by the commutator. The new direction of the current in the armature coil is such as to continue to produce counterclockwise torque. As the rotor rotates in a counterclockwise direction, a voltage proportional to the speed is generated in the armature coil (see *Direct-current generators* above). While this coil voltage is alternating, the commutator action produces a unidirectional voltage at the motor terminals with the polarity shown. The electrical input will be the product of this terminal voltage and the input current. The mechanical output power will be the product of the rotor torque and speed.

In a practical DC motor, the armature winding consists of a number of coils in slots, each spanning $1/p$ of the rotor periphery for p poles. In small motors the number of coils may be as low as six, while in large motors it may be as large as 300. The coils are all connected in series, and each junction is connected to a commutator bar, as indicated in Figure 63. If current enters at the positive brush, the coil currents have the directions shown. All coils under the poles contribute to torque production.

The motor in Figure 63 contains two poles made of ferrite permanent-magnet material. This structure is typical of small DC motors such as those used in automobile fans. When higher torque is required, as, for example, in

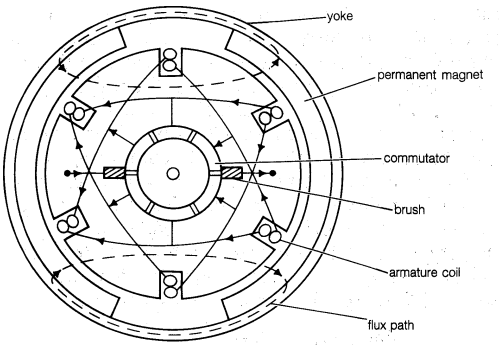


Figure 63: Direct-current commutator motor with a permanent-magnet field.

the starter motor of an automobile, stronger magnets such as neodymium-iron-boron may be employed. When the terminals of this motor are connected to a constant direct-voltage source, such as a battery, the initial current will be limited only by the resistance of the armature winding and the brushes. The torque produced by the interaction of this current with the field accelerates the rotor. A voltage is generated in the winding proportional to the speed. This voltage opposes that of the source, thus reducing the current and the torque. With no mechanical load, the generated voltage will rise to a value nearly equal to the source voltage, allowing just enough current to provide for friction torque. Application of a load torque slows down the rotor, decreasing the generated voltage, increasing the current, and producing torque to match the load torque.

With larger motors, the armature winding resistance is too low to limit the current on starting to a value that can be switched by the commutator. These motors are normally started with a resistance connected in series to the armature supply. This resistance is usually decreased in stages as the speed increases.

Speed-adjustment capability

The permanent-magnet commutator motor of Figure 63 has no provision for speed control when attached to a constant-voltage supply. If speed adjustment is desired, the permanent-magnet field can be replaced by iron poles with field coils. These coils can be provided with current from the same supply as for the armature or from a separate supply. A variable series resistor can be used to adjust the field current. With maximum field current and thus maximum magnetic flux, the generated voltage will equal the supply voltage at a minimum value of no-load speed. As load is added, the speed will reduce somewhat and the armature current will increase to produce the required torque. If the field current is reduced, the motor will have to rotate faster through the reduced flux to generate the same voltage. The no-load speed will be increased. For a given rated armature current, the available torque will be reduced because of the reduced flux. The motor, however, will be able to provide the same mechanical power at a higher speed and lower torque.

Shunt motors

Commutator motors with adjustable field current are known as shunt motors, or separately excited motors. Normally, the available speed range is less than 2 to 1, but special motors can provide a speed range of up to 10 to 1.

Another form of commutator motor is the series motor in which the field coils, with relatively few turns, carry the same current as does the armature. With a high value of current, the flux is high, making the torque high and the speed low. As the current is reduced, the torque is reduced and the speed increases. In the past, such motors were widely used in electric transportation vehicles, such as subway trains and fork-lift trucks.

Large DC motors usually have four or more poles to reduce the thickness of the required iron in the stator yoke and to reduce the length of the end connections on the armature coils. These motors may also have additional small poles, or interpoles, placed between the main poles and have coils carrying the supply current. These poles are placed so as to generate a small voltage in each armature coil as it is shorted out by the commutator. This assists the quick reversal of current in the coil and prevents commutator sparking.

DC commutator motors have been extensively used in steel mills, paper mills, robots, and machine tools where accurate control of speed or speed reversal, or both, are required. The field is supplied from a separate voltage source, usually with constant field current, or from permanent magnets. The armature is supplied from a source of controllable voltage. The speed is then approximately proportional to the source voltage. Reversal of the armature supply voltage at a controlled rate reverses the motor.

Alternating-current commutator motors. A specially designed series-commutator motor may be operated from a single-phase alternating voltage supply. When the supply current reverses, both the magnetic field and the armature current are reversed. Thus, the torque remains in the same direction. These motors are often called universal motors because they may be used with either a direct-voltage supply or with a 60-hertz alternating-voltage supply. They have wide application in such small domestic appliances as mixers, portable tools, and vacuum cleaners.

Universal motors

DEVELOPMENT OF ELECTRIC GENERATORS AND MOTORS

Within a year of Michael Faraday's discovery of electromagnetic induction (1831), a small hand generator was demonstrated in Paris, and by 1850 generators were being manufactured in several countries. These early generators were little more than assemblies of coils and permanent magnets that could be maintained in relative motion. Further developments of significance did not appear until the experimental work of William Sturgeon of England and of Joseph Henry and Thomas Davenport of the United States led to the manufacture of practical electromagnets. This technological advance contributed much to the development of practical electrical machines.

First electric motor of commercial significance

The French engineer and inventor Zénobe-Théophile Gramme built the first truly commercial electric motor, which he demonstrated in 1873. Using iron-cored electromagnets and an iron ring armature surrounded by a winding, Gramme produced a practical, efficient machine that could be used either as a motor or as a generator. His machine was of the DC commutator type. It provided the basis for early DC electric supply.

The first significant AC motor was patented by the Serbian-American inventor Nikola Tesla in 1888. Tesla's motor was able to utilize the two- and three-phase alternating-current supplies that were becoming readily available at the time. Its principle of operation provides the basis for the majority of electric motors produced today. (G.R.Sl.)

Direct energy-conversion devices

BATTERIES

General characteristics. A battery is a simple device that converts chemical energy directly to electrical energy. It consists of two or more galvanic, or electrochemical, cells that produce direct-current electricity. The term battery is also commonly applied to a single galvanic cell. Every battery (or cell) has a cathode, or positive electrode, and an anode, or negative electrode. These electrodes must be separated by and are often immersed in an electrolyte that permits the passage of ions between the electrodes (Figure 64). The electrode materials and electrolyte are chosen and arranged so that sufficient electromotive force (voltage) and electric current (amperes) can be developed between the terminals of a battery to operate lights, machines, or other devices. Since an electrode contains only a limited number of units of chemical energy convertible to electrical energy, it follows that a battery of a given size has a certain capacity to operate devices and will eventually become exhausted. The active parts of a battery are usually encased in a box (or jacket) and cover system that keeps air outside and the electrolyte solvent inside and that provides a structure for the assembly.

Factors that affect battery performance

Battery usefulness is limited not only by capacity but also by how fast current can be drawn from it. The salt ions chosen for the electrolyte solution must be able to move fast enough through the solvent to carry chemical matter between the electrodes equal to the rate of electrical demand. Battery performance is thus limited by the diffusion rates of internal chemicals as well as by capacity.

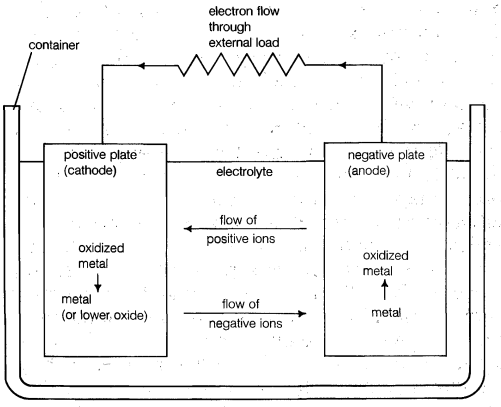


Figure 64: Basic components of an electrochemical cell.

The voltage of an individual cell and the diffusion rates inside it are both reduced if the temperature is lowered from a reference point, such as 21° C. If the temperature falls below the freezing point of the electrolyte, the cell will usually produce very little useful current and may actually change internal dimensions, resulting in internal damage and diminished performance even after it has warmed up again. If the temperature is raised deliberately, faster discharge can be sustained, but this is not generally advisable because the battery chemicals may evaporate or react spontaneously with one another, leading to early failure.

Beyond the technical factors so far discussed, it must be recognized that commercially available batteries are designed and built with market factors in mind. The quality of materials and the complexity of electrode and container design are reflected in the market price sought for any specific product. As new materials are discovered or the properties of traditional ones improved, however, the typical performance of even older battery systems sometimes increases by large percentages.

Batteries are divided into two general groups: (1) primary batteries and (2) secondary, or storage, batteries. Primary batteries are designed to be used until the voltage is too low to operate a given device and then discarded. Secondary batteries have many special design features, as well as particular materials for the electrodes, that permit them to be reconstituted (recycled). After partial or complete discharge, they can be recharged by DC voltage and current to their original state. While this original state is usually not restored completely, the loss per cycle in commercial batteries is only a small fraction of 1 percent even under varied conditions.

Principles of operation. The anode of an electrochemical cell (Figure 64) is usually a metal that is oxidized (gives up electrons) at a potential between 0.5 volt and about four volts above that of the cathode. The cathode generally consists of a metal oxide or sulfide that is converted to a less-oxidized state by accepting electrons, along with ions, into its structure. A conductive link via an external circuit (*e.g.*, a lamp or other device) must be provided to carry electrons from the anode to the negative battery contact. Sufficient electrolyte must be present as well. The electrolyte consists of a solvent (water, an organic liquid, or even a solid) and one or more chemicals that dissociate into ions in the solvent. These ions serve to deliver electrons and chemical matter through the cell interior to balance the flow of electric current outside the cell during cell operation.

Anode and cathode composition

The fundamental relationship of electrochemical cell operation put forth by Faraday in 1834 is that for every ampere that flows for a period of time a matching chemical reaction or other change must take place. The extent of these changes is dependent on the molecular and electronic structure of the elements comprising the battery electrodes and electrolyte. Secondary changes may also occur, but a primary pair of theoretically reversible reactions must take place at the electrodes for electricity to be produced. The actual DC power generated by a battery is measured by the number of amperes produced × the unit of time × the average voltage over that time. For a cell with electrodes of zinc and manganese dioxide (*e.g.*, the common flashlight dry cell), one finds that a chemical equivalent of zinc weighs 32.5 grams and that of manganese dioxide about 87 grams. The discharge of one equivalent weight of each of these electrodes will cause 32.5 grams of zinc to dissolve and 87 grams of manganese dioxide to change into a different oxide containing more hydrogen and zinc ions. Some of the electrolyte also will be consumed in the reaction. One chemical equivalent of each electrode produces one faraday, or 96,500 coulombs of current equal to 26.8 amperes per hour. If the cells operate at an average of 1.2 volts, this would yield 32.2 watt-hours of DC energy. Expressed another way,

$$\text{Energy (joules)} = nFV,$$

where n equals the number of chemical equivalents discharged, F is the Faraday constant (9.648×10^4 coulombs per mole), and V is the average (not necessarily constant) voltage of the cell for the period of the discharge.

There is a large number of elements and compounds from which to select potentially useful combinations for batteries. The commercial systems in common use represent the survivors of numerous tests where continued use depended on adequate voltage, high current-carrying capacity, low-cost materials, and tolerance for user neglect. Better sealing technology and plastics are making further development of all cell systems possible, but particularly those using very active lithium for the anode. This situation has yielded commercial cells with as much as 3.6 volts on load and very high current-carrying capability.

Primary batteries. *Zinc–manganese dioxide systems.* These cell systems are the most commonly used worldwide in flashlights, toys, radios, tape recorders, and flash cameras. There are three variations: the Leclanché cell, the zinc chloride cell, and the alkaline cell. All provide an initial voltage of 1.58 to 1.7 volts, which declines with use to an end point of about 0.8 volt. The Leclanché cell (Figure 65) is the least expensive, traditional general-purpose dry cell available nearly everywhere. Invented by the French engineer Georges Leclanché in 1866, it immediately became a commercial success in large sizes because of its readily available low-cost constituent materials. The anode of this primary cell is a zinc alloy sheet or "cup," the alloy containing small amounts of lead, cadmium, and mercury. The electrolyte consists of a saturated aqueous solution of ammonium chloride containing roughly 20 percent zinc chloride. The cathode is made of impure manganese dioxide (usually mined from selected deposits in Africa, Brazil, or Mexico). This compound is blended with carbon black and electrolyte to create a damp, active cathode mixture which is formed around a carbon collector rod, also called an electrode. All cells of this type are provided with an overwrap structure with metal covers for electrical contact.

Leclanché cell

While first patented in 1899, the zinc chloride cell is really a modern adaptation of the Leclanché cell. Its commercial success is attributable in part to the development of plastic seals that has made it possible to largely dispense with the use of ammonium chloride (Figure 66).

Zinc chloride cell

From G. Vinal, *Storage Batteries* (© 1951); John Wiley & Sons, Inc.

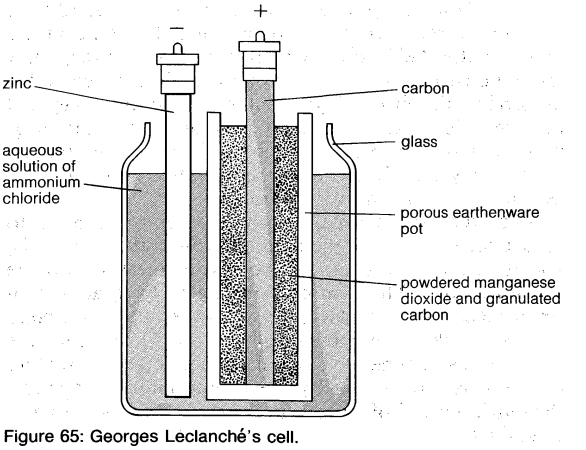


Figure 65: Georges Leclanché's cell.

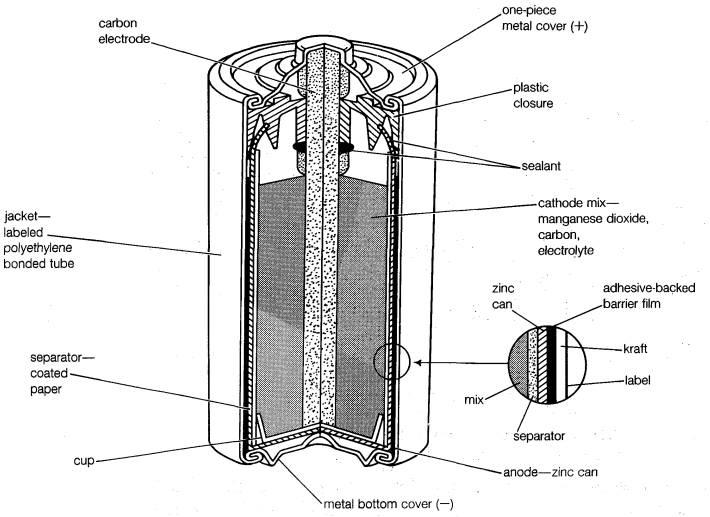


Figure 66: *Modern version of the Leclanché cell.* This heavy-duty carbon–zinc primary battery is a dry cell with an immobilized electrolyte.

By courtesy of Eveready Battery Co., Inc.

The manganese dioxide of the cathode is usually a blend of synthetic manganese dioxide of high purity with natural varieties. The zinc chloride cell is capable of greater continuous service than the Leclanché cell, particularly in motorized devices such as toys. Its use is also increasing because it can provide satisfactory performance without mercury in the zinc alloy.

Alkaline cell

The highest power density (watts per cubic centimetre) of the zinc–manganese dioxide cells is found in cells with an alkaline electrolyte, which permits a completely different type of construction, as illustrated in Figure 67. These cells became commercially available during the 1950s. A cathode of a very pure manganese dioxide–graphite mixture and an anode of a powdered zinc alloy are associated with a potassium hydroxide electrolyte and housed in a steel can. Whereas the zinc of alkaline cells formerly contained 6 to 8 percent mercury, that of present-day versions contains as little as 0.15 percent so as to reduce the environmental impact of disposal. These cells, moreover, provide higher capacity to operate flashlights, toys, cassette players, and radios than either of the other two zinc–manganese dioxide systems discussed above.

Magnesium–manganese dioxide cell. This system functions well for specialized applications. It is much like the zinc chloride cell but has 0.3 volt more per cell. Magnesium–manganese dioxide cells have a long shelf life, high energy density, and are lightweight, making them especially attractive for use as power packs for portable military radios. The one drawback of these cells is that they do not function nearly as well at below-freezing temperatures as at higher temperatures.

By courtesy of Eveready Battery Co., Inc.

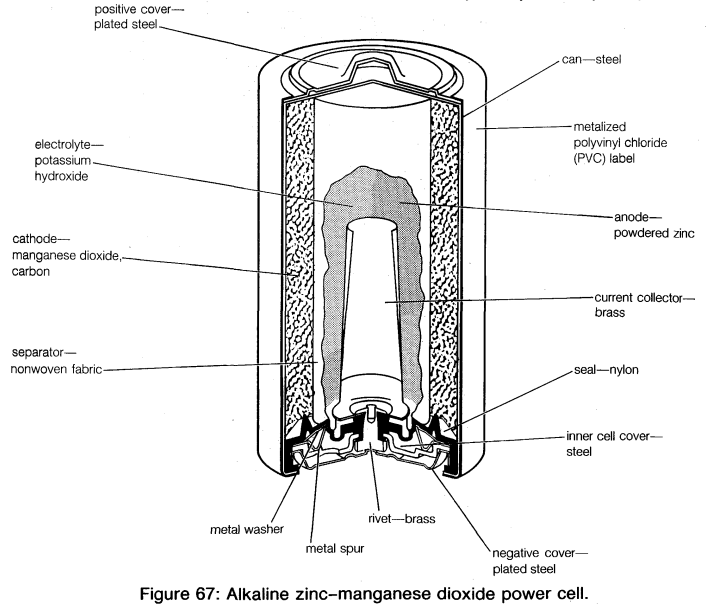


Figure 67: Alkaline zinc–manganese dioxide power cell.

Mercuric oxide–zinc cell. This is an alkaline-electrolyte battery system. It has long been used in the form of button-sized cells (Figure 68) for hearing aids and watches. Its energy density (watt-hours per cubic centimetre) is approximately four times greater than that of the alkaline zinc–manganese dioxide cell. Since the mercuric oxide–zinc cell provides an extremely reliable 1.35 volts, it serves as a standard reference cell.

Silver oxide–zinc cell. Another alkaline system, this cell features a silver oxide cathode and a powdered zinc anode. Because it will tolerate relatively heavy current load pulses and has a high, nearly constant, 1.5-volt operating voltage, the silver oxide–zinc cell is commonly used in watches, cameras, and hearing aids. In spite of its high cost, the outstanding current-carrying capability of this cell has resulted in its use as military torpedo batteries. Miniature cells can be obtained with either divalent silver oxide or monovalent silver oxide, the former usually having somewhat higher capacity.

By courtesy of Eveready Battery Co., Inc.

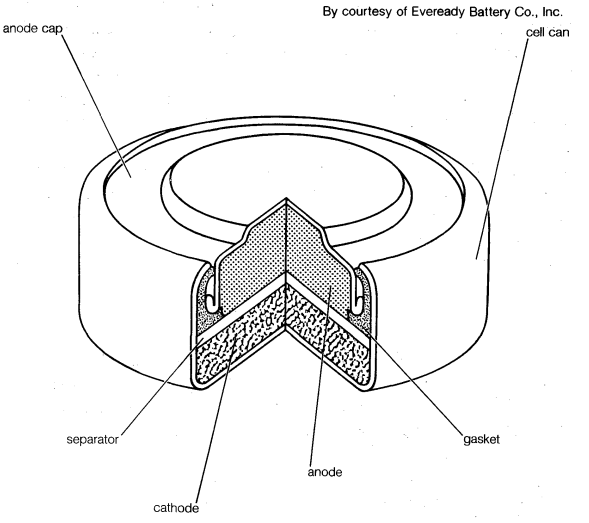


Figure 68: Typical construction of a miniature power cell, as, for example, a silver oxide–zinc or mercuric oxide–zinc system.

Lithium cells. The area of battery technology that has attracted the most research in recent years is a class of cells with a lithium anode. Because of the high chemical activity of lithium, nonaqueous (organic or inorganic) electrolytes have to be used. Such electrolytes include selected solid crystalline salts (see below). This whole new science has encouraged the commercial production of cells having no space between the anode and the liquid cathode, an unlikely condition for success in aqueous systems. A stable protective layer automatically forms on the lithium but breaks down on discharge to permit high-current operation at nearly constant voltages near 3.6 volts. By traditional measures, this allows very high power density and energy density. Lithium cells are especially attractive for use in certain aerospace applications, terrestrial portable military equipment, and such civilian applications as personal paging systems, heart pacers, and automated cameras.

Major types of lithium batteries

Lithium–iron sulfide cells in miniature sizes offer high capacity and low cost for light loads. In operations requiring 1.5 to 1.8 volts, they are a potential substitute for some silver oxide–zinc cells. In constructions where the electrodes consist of rolled up ("jelly roll") strips like those of small nickel–cadmium cells, higher power density is obtained while still retaining high capacity for premium general-purpose use. A typical electrolyte might be lithium tetrafluoborate salt in a solvent mixture of propylene carbonate, 2-methyl-2-oxazolidone, and dimethoxyethane.

Lithium–manganese dioxide cell systems have slowly gained increasingly wider application in small appliances.

Cells of this kind have an operating voltage of 2.8 volts each and offer high energy density and relatively low cost compared to some other lithium cell possibilities.

The lithium–carbon monofluoride system has been among the more successful early commercial lithium cells. It has been used extensively in cameras and smaller devices, providing about 3.2 volts per cell, high power density, and long shelf life. Good low-temperature performance and a flat voltage–time discharge relationship are provided as well. The cost of carbon monofluoride (CF_x) is high, however.

Lithium–thionyl chloride cells provide the highest energy density and power density commercially available. Thionyl chloride serves not only as the electrolyte solvent but also as the cathode material. A runaway reaction between the lithium anode and the adjacent liquid cathode material is prevented by the formation of a film of lithium chloride salt on the lithium. The electrical contact and reaction centre of the cathode are composed of porous pressed and bonded carbon powder. The performance of this type of cell system at room temperature is very impressive. Moreover, the cell can operate at −54° C, well below the point where aqueous systems function. Because of its high energy density, the lithium–thionyl chloride cell must be used with care and not be burned or disposed of casually. Such cells are useful for powering military equipment, providing backup power for aerospace systems, and operating personal pagers.

Lithium–sulfur dioxide cells have been used extensively for some emergency-aircraft power units and in military cold-weather applications (*e.g.*, radio operation). The cathode consists of a gas under pressure with another chemical as electrolyte salt; this is analogous to the thionyl chloride electrolyte and its liquid cathode. The system functions well but has been found to occasionally vent noxious sulfur dioxide, especially after cold discharge and subsequent warm-up. The release of corrosive or toxic gases by any type of cell in a closed space constitutes a significant design disadvantage.

Air-depolarized cells. A very practical way to obtain high energy density in a cell is to employ the oxygen in air for a "liquid" cathode material. If paired with an anode such as zinc, long cell life at low cost per watt-hour (for a dry cell) can be obtained because a given cell volume may be devoted more completely to anode and electrolyte material. The cell, however, must be constructed in such a way that the oxygen is prevented from reaching the anode, which it will attack.

Zinc–air cells

Zinc–air systems are commercially available in the form of very small cells and relatively large boxlike batteries. Their principle and design are simple, but the actual batteries are, from a technical standpoint, difficult to manufacture. The "air electrode" is extremely thin and usually has a waterproof polymer-bonded porous carbon layer with a metal mesh reinforcement. A catalyst and a booster oxide may be included with the carbon to render oxygen more effectively active. The sealing of the edges of the composite electrode film and electrolyte proofing of the pores have been achieved with fluorocarbons and plastics. Fundamental improvements in electrode assembly, cell seal, and vent designs continue to be sought in scientific and engineering studies.

Aluminum–air cells have not been a commercial success to date, but their light weight and potentially high energy density have attracted much government support in the United States. Research efforts have been concentrated on developing better aluminum alloys and techniques to resist corrosion during shelf storage while at the same time providing electricity at instant demand. Similarly, inhibitors for inclusion in the alkaline electrolyte are under study. Aluminum–air cells also are being considered for applications in which the metal anode, the electrolyte, and the reaction products are mechanically removed and replaced to create a kind of fuel cell (see *Fuel cells* below). If stability and cell design problems can be overcome, this system may very well prove attractive for many applications, including use in electric cars or trucks.

Other primary battery systems. Many other cell types are in use on a small scale. For example, cells that produce a very predictable standard voltage are the Clark cell (zinc–mercurous sulfate–mercury; 1.434 volts) and the Weston cell (cadmium–mercurous sulfate–mercury; 1.019 volts). For the construction of the latter, see Figure 69. Magnesium–silver chloride and magnesium–lead chloride batteries are commonly employed in undersea operations where the salt water becomes the electrolyte when the battery is submerged.

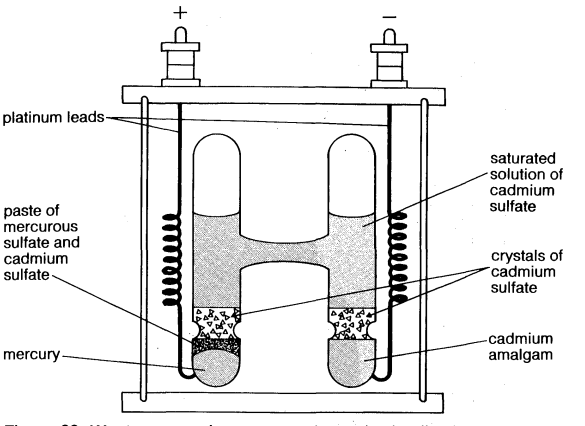


Figure 69: Weston normal or saturated standard cell.
From G. Vinal, *Storage Batteries* (© 1951), John Wiley & Sons, Inc.

An important new group of cells consists of systems with a solid electrolyte in which the mixture of compounds is such that cell ions can slowly move from site to site in the electrolyte crystal structure. Examples include silver–silver rubidium iodide–iodine cells and lithium–lithium iodide–lead iodide mixtures. Batteries with ion-containing polymers are being studied extensively. In such devices, electrode conductivity is achieved by special polymer structure and doping with charged ions either chemically or electrically.

Storage batteries. *Lead secondary cells.* The so-called lead-acid secondary battery has long been the most widely used rechargeable portable power source. Most such batteries are constructed of lead plates, or grids, where one of the grids, the positive electrode, is coated with lead dioxide in a particular crystalline form, along with additives such as calcium lignosulfate (Figure 70). The electrolyte, composed of sulfuric acid, participates in the electrode reactions where lead sulfate is formed and carries current in moving ions. Recent estimates show that in terms of

Components of lead-acid batteries

From G. Vinal, *Storage Batteries* (© 1951); John Wiley & Sons, Inc.

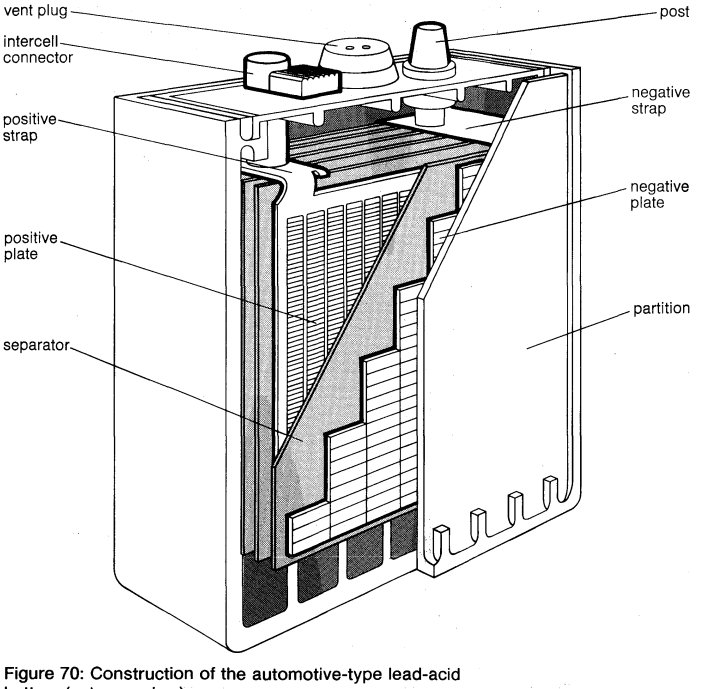


Figure 70: Construction of the automotive-type lead-acid battery (cutaway view).

capacity in use (watt-hours), the lead-acid battery has 20 times as much capacity as either the nickel–cadmium or nickel–iron alkaline rechargeable battery.

The lead-acid battery system has been as successful as it has because of the following features: wide capability range for high or low current demand over usual ambient temperatures; good cycle life with high reliability for hundreds of cycles, especially with good recharge control (a gram of positive active material may deliver as many as 100 ampere-hours during the service life of such a battery); relative low cost (lead is less expensive per kilogram or per ampere-hour than nickel, cadmium, or silver); comparatively good shelf life for a rechargeable system when stored; high cell voltage at 2.04 volts per cell; ease of fabricating lead components by casting, welding, or rolling; and a high degree of salvageability at low melting temperatures.

An area of continued interest for investigators working on lead-acid batteries is reduction of battery weight. Lead dioxide and lead have the lowest energy density of the major electrode materials in wide use, and they are rarely discharged in a highly efficient manner. At low rates of discharge, only about 60 percent of the active materials are cycled, and on short, 10-minute heavy loads utilization can fall to 10 percent.

Types of lead-acid batteries

Lead-acid batteries are generally classified into three groups: (1) starting–lighting–ignition (SLI) batteries, (2) traction batteries, and (3) stationary batteries. The automotive SLI battery is the best known portable rechargeable power source. High current can be obtained for hundreds of shallow-depth discharges over a period of several years. Traction batteries are employed in industrial lift trucks, delivery trucks, and other vehicles. While some are readily portable, others may weigh several tons. The great weight often serves to stabilize the vehicle during operation. Stationary batteries are now much more common than was once the case. These batteries have heavier grid structures and other features to give them long shelf life. They are used to power emergency lights and in uninterruptible power systems for hospitals, factories, and telephone exchanges.

In a lead-acid battery the active material of the positive electrode, lead dioxide, combines with the electrolyte, sulfuric acid, to produce lead sulfate and water during discharge. At the negative electrode the constituent lead combines with the sulfuric acid ions to produce lead sulfate and hydrogen ions, thereby replacing the hydrogen ions consumed at the positive electrode. The water formed and the loss of sulfate dilutes the electrolyte, lowering its density. Because of this, the state of charge of a lead-acid battery can be determined from the specific gravity of the electrolyte.

Alkaline storage batteries. In secondary batteries of this type, electric energy is derived from the chemical action in an alkaline solution. Such batteries feature a variety of electrode materials, some of the more notable of which are briefly discussed in this section.

Nickel–cadmium batteries

Nickel (hydroxide)–cadmium systems are the most common small rechargeable battery type for portable appliances. The sealed cells are equipped with "jelly roll" electrodes (see Figure 71), which allow high current to be delivered in an efficient way. These batteries are capable of delivering exceptionally high currents, can be rapidly recharged hundreds of times, and are tolerant of abuse such as overdischarging or overcharging. Nonetheless, compared to many primary batteries and even lead-acid batteries, nickel–cadmium cells are heavy and have comparatively limited energy density. They last longer and perform better if fully discharged each cycle before recharge. Otherwise, the cells may exhibit a so-called memory effect where they behave as if they had lower capacity than was built into the battery pack. Larger nickel–cadmium batteries are used for starting up aircraft engines and in emergency power systems. They also have found application in other backup power systems where very high currents, low temperature conditions, and reliability are special factors. In addition, they are used in tandem with a solar-powered current source to provide electric power at night.

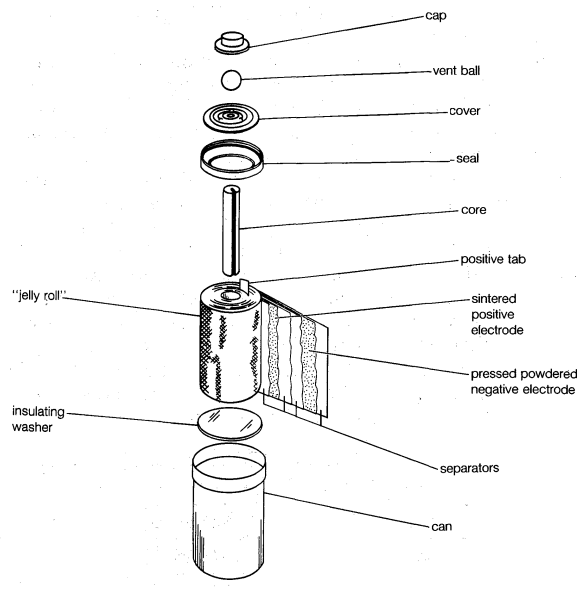


Figure 71: Nickel (hydroxide)–cadmium cell of "jelly roll" construction (see text).
By courtesy of Eveready Battery Co., Inc.

Nickel (hydroxide)–zinc cells are attractive from a development viewpoint. If their cycle life can be significantly improved, systems of this sort may become a viable substitute for nickel–cadmium cells or lead-acid traction batteries.

Nickel (hydroxide)–iron batteries can provide thousands of cycles but do not recharge with high efficiency, generating heat and consuming more electricity than is generally desirable. They have been used extensively in the European mining industry, however.

Nickel–hydrogen batteries

Nickel (hydroxide)–hydrogen cells were developed primarily for the U.S. space program. Research has shown that such alloys as lanthanum-nickel in certain proportions will reversibly dissolve or release hydrogen in proportion to changes in pressure and temperature. This hydrogen can serve as an active anode material. There is speculation that nickel–hydrogen batteries may replace nickel–cadmium batteries in many applications.

Alkaline zinc–manganese dioxide rechargeable cells have been developed as a substitute for other systems that provide moderate amounts of electricity for certain applications. Their high energy density and low cost encourage further engineering work and commercial introduction.

Silver (oxide)–zinc batteries are expensive but are employed where high power density, good energy-cycling efficiency, low weight, and low volume are critical. After years of use in torpedoes and mines, they have more recently become important in special vehicles for underwater tests and submarine exploration. They also are employed in portable radar units and communications equipment, as well as in aircraft and space vehicles.

Lithium secondary cells. These show considerable promise since their theoretical energy densities can range from 600 to 2,000 watt-hours per kilogram. Even after allowance is made for the inactive parts of such a cell, the net energy density is still competitive with aqueous systems. Systems of this type receiving developmental attention include lithium–titanium disulfide, lithium–manganese dioxide, and lithium–molybdenum disulfide. Much current research is devoted to developing better oxide and sulfide structures, better solvent combinations, and better and safer constructions.

Sodium–sulfur storage batteries. Much experimental work has been expended on this type of high-temperature system, which operates at 350° C. Many problems related to material stability have to be solved before a completely satisfactory system can be produced. This is particularly true given the need to tolerate cooling and heating the whole battery between uses. Yet, the ready availability of sodium and sulfur, low cost, and ability of each cell to deliver 2.3 volts make this system extremely attractive. It

could be used for electric vehicles or to help meet municipal peak power requirements.

Development of batteries. The Italian physicist Alessandro Volta is generally credited with having developed the first operable battery. Following up on the earlier work of his compatriot Luigi Galvani, Volta performed a series of experiments on electrochemical phenomena during the 1790s (see ELECTRICITY AND MAGNETISM). By about 1800 he had built his simple battery, which later came to be known as the "voltaic pile." This device consisted of alternating zinc and silver disks separated by layers of paper or cloth soaked in a solution of either sodium hydroxide or brine (Figure 72). Experiments performed with the voltaic pile eventually led Faraday to derive the quantitative laws of electrochemistry (about 1834). These laws, which established the exact relationship between the quantity of electrode material and the amount of electric power desired, formed the basis of modern battery technology.

The voltaic pile

From G. Vinal, *Primary Batteries* (© 1950), John Wiley & Sons, Inc., after the *Philosophical Transactions of the Royal Society* (1800)

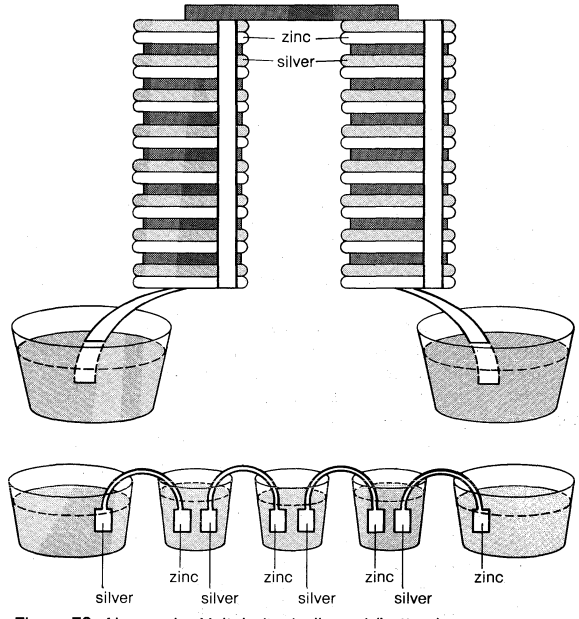


Figure 72: Alessandro Volta's (top) pile and (bottom) crown of cups.

Various commercially significant primary cells were produced on the heels of Faraday's theoretical contribution. In 1836 John Frederic Daniell, a British chemist, introduced an improved form of electric cell consisting of copper and zinc in sulfuric acid. The Daniell cell was able to deliver sustained currents during continuous operation far more efficiently than Volta's device.

Further advances were effected in 1839 by William Robert Grove with his two-fluid primary cell consisting of amalgamated zinc immersed in dilute sulfuric acid, with a porous pot separating the sulfuric acid from a strong nitric acid solution containing a platinum cathode. The nitric acid served as an oxidizing agent, which prevented voltage loss resulting from an accumulation of hydrogen at the cathode. The German chemist Robert Wilhelm Bunsen substituted inexpensive carbon for platinum in Grove's cell and thereby helped promote its wide acceptance.

In 1859 Gaston Planté of France invented a lead-acid cell, the first practical storage battery and the forerunner of the modern automobile battery. Planté's device was able to produce a remarkably large current, but it remained a laboratory curiosity for nearly two decades.

First practical storage battery

Georges Leclanché's prototype of the zinc–manganese dioxide system paved the way for the development of the modern primary cell. The original version of the Leclanché cell was "wet," as it had an electrolyte consisting of a solution of ammonium chloride. The idea of employing an immobilized electrolyte was finally introduced in the late 1880s and launched the dry-cell industry that continues to flourish today.

The invention of alkaline electrolyte batteries (specifically storage batteries of the nickel–cadmium and nickel–iron type) between 1895 and 1905 provided systems that could furnish much-improved cycle life for commercial application. The 1930s and '40s saw the development of the silver oxide–zinc and mercuric oxide–zinc alkaline cells, systems that provided the highest energy yet known per unit weight and volume. Since midcentury, advances in construction technology and the availability of new materials have given rise to smaller yet more powerful batteries suitable for use in a wide array of portable equipment. Perhaps most notable have been the entrance of lithium batteries into the commercial market and the development of nickel–hydrogen cells for use in spacecraft.

FUEL CELLS

General characteristics. A fuel cell is an electrochemical device that converts the chemical energy of a fuel directly and efficiently to direct-current electricity in a continuous manner. It resembles a battery in many respects, but it can supply electrical energy over a much longer period of time. This is because a fuel cell is continuously supplied with fuel and air (or oxygen) from an external source, while a battery contains only a limited amount of fuel material and oxidant, which are depleted with use.

External source of fuel and oxidizer

A fuel cell (actually a group of cells) has essentially the same kinds of components as a battery. As in the latter, each cell of a fuel-cell system has a matching pair of electrodes (Figure 73). These are the anode, which supplies electrons, and the cathode, which absorbs electrons. Both electrodes must be immersed in and separated by an electrolyte, which may be a liquid or a solid but which must in either case conduct ions between the electrodes in order to complete the chemistry of the system. A fuel, such as hydrogen, is supplied to the anode where it is oxidized, producing hydrogen ions and electrons. An oxidizer, such as oxygen, is supplied to the cathode where the hydrogen ions from the anode absorb electrons from the latter and react with the oxygen to produce water. The difference between the respective energy levels at the electrodes (electromotive force) is the voltage per unit cell. The amount of current available to the external circuit depends on the chemical activity and amount of the substances supplied as fuels. The current-producing process continues for as long as there is a supply of reactants, for, unlike in a regular battery, the electrodes and electrolyte of a fuel cell are designed to remain unchanged by chemical reaction.

A practical fuel cell is necessarily a complex system. It must have features to boost the activity of the fuel, pumps and blowers, fuel-storage containers, and a variety of sophisticated sensors and controls with which to monitor and adjust the operation of the system. The operating capability and lifetime of each of these system design features may limit the performance of the fuel cell.

As in the case of other electrochemical systems, fuel-cell operation is dependent on temperature. The chemical activity of the fuels and the value of the activity promoters, catalysts, are reduced by low temperatures (*e.g.*, 0° C). Very high temperatures, on the other hand, improve the

Temperature-dependent operation

From R. Noyes (ed.), *Fuel Cells for Public Utility and Industrial Power* (1977), Noyes Data Corp., based on a report by United Technologies Corp. (February 1976)

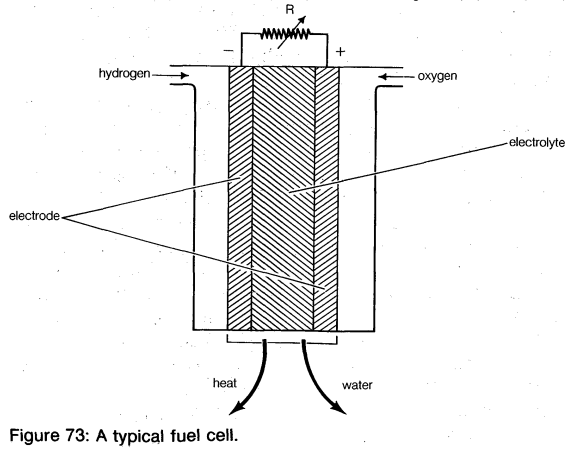


Figure 73: A typical fuel cell.

activity factors but may reduce the functioning lifetime of electrodes, blowers, construction materials, and sensors. Each type of fuel cell thus has an operating-temperature design range, and a significant departure from this range is likely to diminish both capacity and lifetime.

A fuel cell, like a battery, is inherently a high-efficiency device. Unlike internal-combustion machines, where a fuel is burned and gas is expanded to do work, the fuel cell converts chemical energy directly into electrical energy (Figure 74). Because of this fundamental characteristic, fuel cells may convert fuels to useful energy at an efficiency as high as 60 percent, whereas the internal-combustion engine is limited to efficiencies of near 40 percent or less. The high efficiency means that much less fuel and a smaller storage container are needed for a fixed energy requirement. For this reason, fuel cells are an attractive power supply for space missions of limited duration and for other situations where fuel is very expensive and difficult to supply. They also emit no noxious gases such as nitrogen dioxide and produce virtually no noise during operation, making them contenders for local municipal power generation stations.

From R. Noyes (ed.), *Fuel Cells for Public Utility and Industrial Power* (1977), Noyes Data Corp., based on a report by United Technologies Corp. (February 1976)

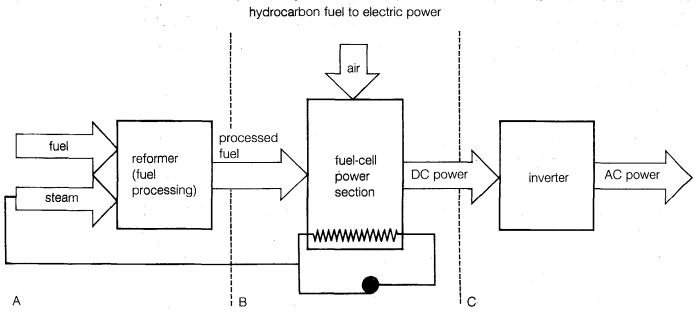


Figure 74: *Elements of a fuel-cell power plant.* (A) The reformer section processes hydrocarbon fuel for fuel-cell use. (B) The power section converts the processed fuel and air into DC power. (C) The inverter produces usable AC power to meet customer requirements.

Regenerative fuel cells

A fuel cell can be designed to operate reversibly. In other words, a hydrogen–oxygen cell that produces water as a product can be made to regenerate hydrogen and oxygen. Such a regenerative fuel cell entails not only a revision of electrode design but also the introduction of special means for separating the product gases. Eventually power modules comprised of this type of high-efficiency fuel cell, used in conjunction with large arrays of solar thermal collectors or other solar power systems, may be utilized to keep energy-cycle costs lower in longer-lived equipment.

Principles of operation. Because a fuel cell produces electricity continuously from fuel, it has many output characteristics similar to those of any other direct-current generator system. A DC generator system can be operated in either of two ways from a planning viewpoint: (1) Fuel may be burned in a heat engine to drive an electric generator, which makes power available and current flow. Or (2) fuel may be converted to a form suitable for a fuel cell, which then generates power directly.

Available fuels

A wide range of liquid and solid fuels may be used for a heat-engine system, while hydrogen, reformed natural gas (*i.e.,* methane that has been converted to hydrogen-rich gas), and methanol are the primary fuels available for current fuel cells. If fuels such as natural gas must be altered in composition for a fuel cell, the net efficiency of the fuel-cell system is reduced, and much of its efficiency advantage is lost. Such an "indirect" fuel-cell system would still display an efficiency advantage of as much as 20 percent. Nonetheless, to be competitive with modern thermal generating plants a fuel-cell system must attain a good design balance with low internal electrical losses, corrosion-resistant electrodes, electrolyte of constant composition, low catalyst costs, and ecologically acceptable fuels.

The first technical challenge that must be overcome in developing practical fuel cells is to design and assemble consistently an electrode that allows the gaseous or liquid fuel to contact a catalyst and an electrolyte at a group of solid sites that do not change very rapidly. Thus, a three-phase reaction situation is typical on an electrode that must also serve as an electrical conductor. As seen in Figure 75, such can be provided by thin sheets that have (1) a waterproof layer usually with Teflon (polytetrafluoroethylene), (2) an active layer of a catalyst (*e.g.,* platinum, gold, or a complex organometallic compound on a carbon base), and (3) a conducting layer to carry the current generated in or out of the electrode. If the electrode floods with electrolyte, the operation rate would become very slow at best. If the fuel breaks through to the electrolyte side of the electrode, the electrolyte compartment might become filled with gas or vapour, inviting an explosion should the oxidizing gas also reach the electrolyte compartment or the fuel gas enter the oxidizing gas compartment. In short, careful design, construction, and pressure control are essential in a working fuel cell to maintain stable operation. Since fuel cells have been used on Apollo lunar flights as well as on all other U.S. orbital manned space missions (*e.g.,* those of Gemini and the Space Shuttle), it is evident that all three requirements can be met reliably.

Providing a fuel-cell support system of pumps, blowers, sensors, and controls for maintaining fuel rates, electric current load, gas and liquid pressures, and fuel-cell temperature remains a major engineering design challenge. Significant improvements in the service life of these components under adverse conditions would contribute to the wider use of fuel cells.

Types of fuel cells. Various types of fuel cells have been developed. They are generally classified on the basis of the electrolyte used because the electrolyte determines the operating temperature of a system and in part the kind of fuel that can be employed.

Alkaline fuel cells. These are devices that, by definition, have an aqueous solution of sodium hydroxide or potassium hydroxide as the electrolyte. The fuel is almost always hydrogen gas, with oxygen or oxygen in air as the oxidizer. The cells generally operate at less than 100° C and are constructed of metal and certain plastics. Electrodes are made of carbon and a metal such as nickel. Product water must be removed from the system as a reaction product, usually by evaporation from the electrolyte either through the electrodes or in a separate evaporator. The operating support system presents a significant design problem. The strong, hot alkaline electrolyte attacks most polymers and tends to readily penetrate structural seams and joints. These problems have been overcome, however, and alkaline fuel cells are used on the U.S. Space Shuttle. Overall efficiencies range from 30 to 80 percent, depending on the fuel, oxidizer, and basis for the calculation.

Use on the U.S. Space Shuttle

Phosphoric acid fuel cells. Such cells have an orthophosphoric acid electrolyte that allows operation up to 200° C. They can use a hydrogen fuel contaminated with carbon dioxide and an oxidizer of air or oxygen. The electrodes consist of catalyzed carbon and are arranged in pairs set back-to-back to create a series generation circuit. The framing structure for this assembly of cells is made of graphite, which markedly raises the cost. The higher temperature and aggressive hot phosphate create struc-

Reprinted with permission from *CHEMTECH*, August 1988, 18(8), p. 494; copyright 1988 American Chemical Society

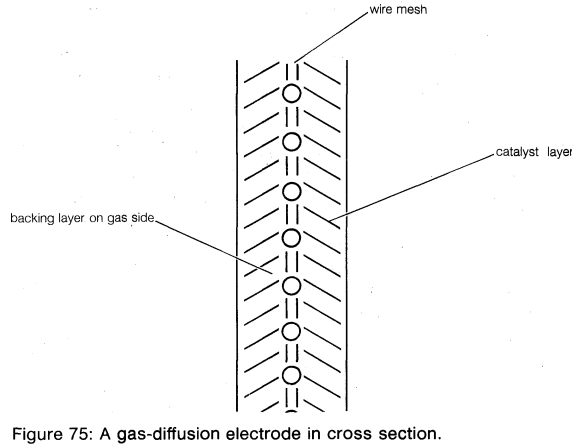


Figure 75: A gas-diffusion electrode in cross section.

tural design problems, particularly for joints, supporting pumps, and sensors. Phosphoric acid fuel cells have been proposed and tested on a limited scale for local municipal power stations and for remote-site generators.

Molten carbonate fuel cells. Fuel cells of this type operate quite differently from those so far discussed. The fuel consists of a mixture of hydrogen and carbon monoxide generated from water and a fossil fuel. The electrolyte is molten potassium lithium carbonate, which permits an operating temperature of about 650° C. In most cases, the electrodes are metallic-based, and the containment system is made of metals and special engineering plastics. (Surprisingly, such combinations of materials are anticipated to be relatively inexpensive, perhaps only three times that of the alkaline fuel cell and less than that of the phosphoric acid variety.) The cells combine the hydrogen and carbon monoxide first with the carbonate electrolyte and then with oxidizing oxygen to produce a reaction product of water vapour and carbon dioxide.

Molten carbonate fuel cells are expected to be useful in both local and larger power stations. Efficiencies of 45 percent may be attained where fossil fuels are already used. Operation at high temperatures creates a design problem for long-lived system parts and joints, especially if the cells must be heated and cooled frequently. The toxic fuel and high temperature together make power-plant safety an area of special concern in engineering design and testing as well as in commercial operation.

Solid oxide fuel cells. In some ways solid oxide fuel cells are similar to molten carbonate devices. Most of the cell materials, however, are special ceramics with some nickel. The electrolyte is an ion-conducting oxide such as zirconia treated with yttria. The fuel for these experimental cells is expected to be hydrogen combined with carbon monoxide, just as for molten carbonate cells. While internal reactions would be different in terms of path, the cell products would be water vapour and carbon dioxide. Because of the high operating temperature (800 to 1,000° C), the electrode reactions proceed very readily. As in the case of the molten carbonate fuel cell, there are many engineering challenges involved in creating a long-lived containment system for cells that operate at such a high-temperature range.

High operating temperature

Solid oxide fuel cells would be designed for use in central power-generation stations where temperature variation could be controlled efficiently and where fossil fuels would be available. The system would in most cases be associated with the so-called bottoming steam (turbine) cycle—*i.e.,* the hot gas product (at 1,000° C) of the fuel cell could be used to generate steam to run a turbine and extract more power from heat energy. Overall efficiencies of 50 to 55 percent might be possible.

Solid polymer electrolyte fuel cells. A cell of this sort is built around an ion-conducting membrane such as Nafion (trademark for a perfluorosulfonic acid membrane). The electrodes are catalyzed carbon, and several construction alignments are feasible. Solid polymer electrolyte cells function well (as attested to by their performance in Gemini spacecraft), but cost estimates are high for the total system compared to the types described above. Engineering or electrode design improvements could change this disadvantage.

Development of fuel cells. The general concept of a fuel battery, or fuel cell, dates back to the early days of electrochemistry. William Grove used hydrogen and oxygen as fuels catalyzed on platinum electrodes in 1839. During the late 1880s two English chemists, Ludwig Mond and Carl Langer, developed a fuel cell with a longer service life by employing a porous nonconductor to hold the electrolyte. It was subsequently found that a carbon base permitted the use of much less platinum, and the German chemist Wilhelm Ostwald proposed as a substitute for heat-engine generators electrochemical cells in which carbon would be oxidized to carbon dioxide by oxygen. During the early years of the 20th century Fritz Haber, Walther H. Nernst, and Edmond Bauer experimented with cells using a solid electrolyte. Limited success and high costs, however, suppressed interest in continuing developmental efforts.

From 1932 until well after World War II, Francis T. Bacon and his coworkers at Cambridge worked on creating practical hydrogen–oxygen fuel cells with an alkaline electrolyte. Research resulted in the invention of gas-diffusion electrodes in which the fuel gas on one side is effectively kept in controlled contact with an aqueous electrolyte on the other side. By mid-century O.K. Davtyan of the Soviet Union had published the results of experimental work on solid electrolytes for high-temperature fuel cells and for both high- and low-temperature alkaline electrolyte hydrogen–oxygen cells.

Invention of gas-diffusion electrodes

The need for high-efficiency stable power supplies for space satellites and manned spacecraft created exciting new opportunities for fuel-cell development during the 1950s and '60s. Molten carbonate cells with magnesium oxide pressed against the electrodes were demonstrated by J.A.A. Ketelaar and G.H.J. Broers of The Netherlands, while the very thin Teflon-bonded, carbon-metal screen catalyzed electrode was devised by other researchers. Many other technological advances, including the development of new materials, played a crucial role in the emergence of today's practical fuel cells. Further improvements in electrode materials and construction, combined with rising fuel costs, are expected to make fuel cells an increasingly attractive alternative power source, especially in Japan and other countries that have meagre nonrenewable energy resources. (B.S.)

SOLAR CELLS

General characteristics. A solar cell is an electronic device that directly converts the energy in light into electrical energy through the process of photovoltaics. Unlike batteries or fuel cells, solar cells do not utilize chemical reactions to produce electric power; and, unlike electric generators, they do not have any moving parts. Solar cells are also called solar batteries and, as the term solar implies, they are in most cases designed for converting sunlight into electrical energy.

Solar cells can be arranged into large groupings called arrays. These arrays, which may be composed of many thousands of individual cells, can function as central electric power stations in the same manner as nuclear power plants and coal- or oil-fired power plants. Such solar-cell power installations convert the energy in sunlight into electrical energy for distribution to industrial, commercial, and residential users. Solar cells in much smaller configurations, commonly referred to as solar-cell panels, are used to provide electric power in many remote terrestrial locations; they are well suited, for example, to run water pumps in desert areas and to power navigational aids at sea. Because they have no moving parts that could require service or fuels that would require replenishment, solar cells are ideal for providing power in space. As a consequence, most space satellites, including communications and weather satellites, are solar-cell powered. Since light is the basic source of the power generated by solar cells, space applications are generally limited to regions of the solar system that are close enough to the Sun to receive substantial amounts of radiant energy. Another growing application of solar cells is in consumer products, such as electronic toys, hand-held calculators, and portable radios. Solar cells used in devices of this kind may utilize indoor artificial light (*e.g.,* from incandescent and fluorescent lamps) as well as natural light from the Sun in converting radiant energy into electricity.

Arrays of solar cells

Structure and principles of operation. The basic structure of a typical solar cell, whether it is used in a central power station, a satellite, or a calculator, is shown in Figure 76. As many be seen, light enters the device through a layer of material called the antireflection layer. The function of this layer is to trap the light falling on the solar cell and to promote the transmission of this light into the energy-conversion layers below. Such materials as silicon oxides or titanium dioxide are employed as the antireflection layer in solar cells. The photovoltaic effect, which causes the cell to convert light directly into electrical energy, occurs in the three energy-conversion layers below the antireflection layer. The first of these three layers necessary for energy conversion in a solar cell is the top junction layer in Figure 76. The next layer in the structure

Energy-conversion layers

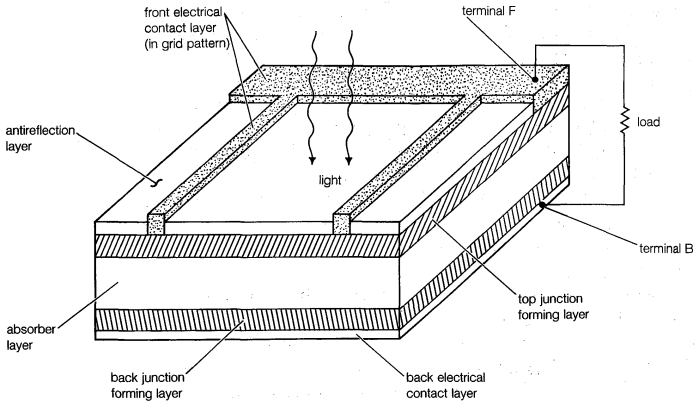


Figure 76: *A commonly used solar-cell structure.* In many such cells, the absorber layer and the back junction layer are both made of the same material.

is the core of the device; this is the absorber layer. The last of the energy-conversion layers is the back junction layer.

As may be seen from Figure 76, there are two additional layers that must be present in a solar cell. These are the electrical contact layers. There must obviously be two such layers to allow electric current to flow out of and into the cell. The electrical contact layer on the face of the cell where light enters is generally present in some grid pattern and is composed of a good conductor such as a metal. The grid pattern does not cover the entire face of the cell since grid materials, though good electrical conductors, are generally not transparent to light. Hence, the grid pattern must be widely spaced to allow light to enter the solar cell but not to the extent that the electrical contact layer will have difficulty collecting the current produced by the cell. The back electrical contact layer has no such diametrically opposed restrictions. It need simply function as an electrical contact and thus covers the entire back surface of the cell structure. Because the back layer must be a very good electrical conductor, it is always made of metal.

It is a fundamental fact of nature that, whenever different materials are placed in contact, an electric field exists at the interface, or junction, between these materials. The role of the junction layers in Figure 76 is to establish this electric field. The field created in the solar cell by the different junction-forming materials is termed the built-in electric field. An electric field is needed in a solar cell because it exerts a force on electrons. If electrons are not attached to specific atoms but are free to roam about in a material, they always will move in a direction dictated by the electric field. This movement constitutes an electric current.

Built-in electric field

The electric field set up by the junction-forming layers of the solar cell causes a current to flow when there are free electrons present in the top junction-forming layer, the absorber layer, and the back junction-forming layer. When light falls on the cell, free electrons occur as a result of the interaction of the light with the absorber layer. The special attribute of this cell layer is that it absorbs light by changing the energy and state (or condition) of some of the electrons in the material. When light is absorbed in the materials, the energy of an electron increases from the so-called ground state energy to an excited energy state. In the excited state, electrons are no longer associated with specific atoms in the absorber, but they are, instead, free to move.

In summary, the absorption of light in the absorber material of a solar cell results in energetic, free electrons that move in the direction forced on them by the built-in electric field. These energetic electrons of the induced current are then collected by the electrical contact layers for use in an external circuit where they can do useful work.

Since most of the energy in sunlight or indoor light is in visible light, a solar-cell absorber should be a strong absorber of electromagnetic radiation in that range of wavelengths. Materials that absorb the visible light of sunlight or of indoor light by producing excited free electrons belong to a class of substances known as semiconductor materials. Semiconductors can absorb all incident visible light in thicknesses of about one-hundredth of a centimetre or less; consequently, the thickness of a solar cell can be of this size. Examples of semiconductor materials employed in solar cells include silicon, gallium arsenide, indium phosphide, and copper indium selenide.

The materials in a solar cell used for the junction-forming layers need only be dissimilar, and, to carry the electric current, they must be conductors. The two junction-forming layers may be different semiconductors or they may be a metal and a semiconductor. Thus, the materials used to construct the various layers of solar cells are essentially the same materials used to produce the diodes and transistors of solid-state electronics and microelectronics (see also ELECTRONICS: *Optoelectronic devices*). Solar cells and microelectronic devices share the same basic technology. In solar-cell fabrication, however, one seeks to construct a large-area device because the power produced is proportional to the illuminated area. In microelectronics the goal is of course to construct devices of very small area to increase the number of circuit components on a single tiny semiconductor chip.

The photovoltaic effect that causes the direct energy conversion in a solar cell is summarized in the schematic of Figure 77. An analogy between an electron in the solar cell and a child at a slide is also presented in this figure. As shown, initially both the electron and the child are in their respective ground states. Next the electron is lifted up to its excited state by consuming energy in the incoming light, just as the child is lifted up to an excited state at the top of the slide by consuming chemical energy stored in his body. In both cases, there is now energy available in the excited state that can be expended. The excited electron is free and moves to the external circuit due to the built-in electric field. It is in this external circuit that the electron will dissipate its excess energy in some device, which in general can be termed a load. The external load is shown here as a simple resistor, but it can be any of a myriad of electrical or electronic devices ranging from motors to radios. Correspondingly, the child moves to the slide because of his desire for excitement. It is on the slide that the child dissipates his excess energy. Finally, when the excess energy is expended, both the electron and the child are back in the ground state where they can, of course, begin the whole process over again. As can be seen from the figure, the motion of the electron, like that of the child, is in one direction. In short, a solar cell produces a direct electric current—namely, one that flows constantly in only a single direction.

Photovoltaic effect

The photovoltaic process bears certain similarities to photosynthesis in plants by which the energy in light is converted into chemical energy. Since solar cells obviously cannot produce electric power in the dark, part of the energy they develop under light is stored, in many applications, for use when light is not available. One common means of storing this electrical energy is to charge chemical

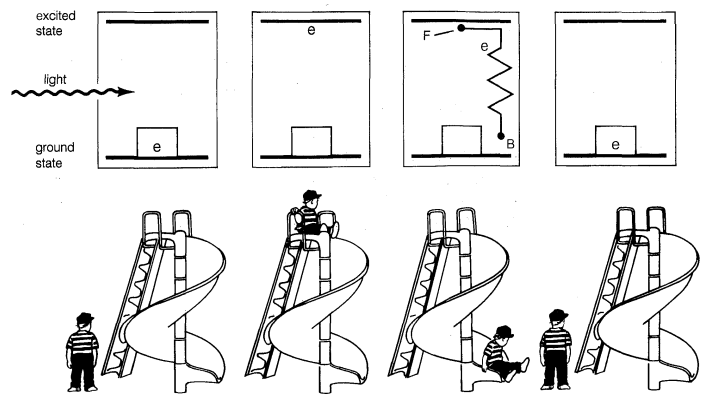


Figure 77: *Representation of an electron in a solar cell.* The electron is shown interacting with light and subsequently dissipating the excess energy it receives from the light by doing work in an external circuit. The electric current flows in and out of the cell through terminals B and F (as represented in Figure 76). The sequence of events involved is analogous to a child playing on a slide (see text).

storage batteries. This sequence of converting the energy in light into the energy of excited electrons and then into stored chemical energy is strikingly similar to the process of photosynthesis.

Development of solar cells. The development of solar-cell technology stems from the work of the French physicist Antoine-César Becquerel in 1839. Becquerel discovered the photovoltaic effect while experimenting with a solid electrode in an electrolyte solution; he observed that voltage developed when light fell upon the electrode. About 50 years later, Charles Fritts constructed the first true solar cells using junctions formed by coating the semiconductor selenium with an ultrathin, nearly transparent layer of gold. Fritts's devices were very inefficient converters of energy; they transformed less than 1 percent of the absorbed light energy into electrical energy. Though inefficient by today's standards, these early solar cells fostered among some a vision of abundant, clean power. In 1891 R. Appleyard wrote of "the blessed vision of the Sun, no longer pouring his energies unrequited into space, but by means of photo-electric cells . . . , these powers gathered into electrical storehouses to the total extinction of steam engines, and the utter repression of smoke."

By 1927 another metal–semiconductor-junction solar cell, in this case made of copper and the semiconductor copper oxide, had been demonstrated. By the 1930s both the selenium cell and the copper oxide cell were being employed in light-sensitive devices, such as photometers, for use in photography. These early solar cells, however, still had energy-conversion efficiencies of less than 1 percent. This impasse was finally overcome with the development of the silicon solar cell by Russell Ohl in 1941. Thirteen years later three other American researchers, G.L. Pearson, Daryl Chapin, and Calvin Fuller, demonstrated a silicon solar cell capable of a 6-percent energy-conversion efficiency when used in direct sunlight. By the late 1980s silicon cells, as well as those made of gallium arsenide, with efficiencies of more than 20 percent had been fabricated. In 1989 a concentrator solar cell, a type of device in which sunlight is concentrated onto the cell surface by means of lenses, achieved an efficiency of 37 percent due to the increased intensity of the collected energy. In general, solar cells of widely varying efficiencies and cost are now available. (S.J.F./R.T.F.)

First silicon solar cell

THERMOELECTRIC POWER GENERATORS

General characteristics. A unique aspect of thermoelectric energy conversion is that the conversion direction is reversible. This distinguishes thermoelectric energy converters from many other energy conversion systems. Electrical input power can be directly converted to pumped thermal power for the purpose of either refrigeration or heating. Conversely, thermal input power can be converted directly to electrical power for lighting, operating electrical equipment, and other work functions. Though any thermoelectric device can be applied in either mode of operation, the design of a particular device may not be optimal.

Capability for reversible energy conversion

All thermoelectric power generators are configured as shown in Figure 78. The heat source provides for the high temperature and the amount of heat flow through the thermoelectric converter to the heat sink. The heat sink is maintained at a temperature below that of the source. The temperature differential, $\Delta T = T_1 - T_0$, across the converter produces direct-current electrical power to a load R (ohms), having a terminal voltage V (volts), and provides a current I (amperes). There is no intermediate conversion process. For this reason, thermoelectric power generation is classified as direct power conversion. The amount of electrical power generated, W (watts), is I^2R_L or alternately VI.

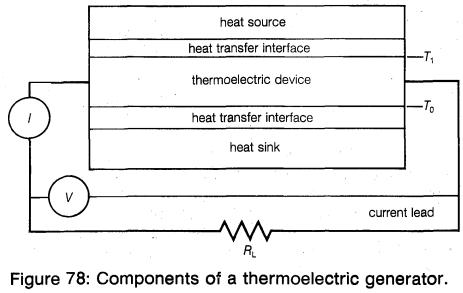


Figure 78: Components of a thermoelectric generator.

If the load resistor is removed and a DC power supply is substituted, the thermoelectric device of Figure 78 can be used to lower the temperature of the heat source, provided that the input thermal power is not increased. In this configuration, the reversed energy-conversion process of thermoelectric devices, using electrical power to pump heat, is invoked.

Principles of operation. An introduction to the phenomenon of thermoelectricity is necessary to understand the operating principles of thermoelectric devices.

In 1821 the German physicist Thomas Johann Seebeck discovered that when two strips of different conductors (metals, semimetals, or semiconductors—the distinction was not understood at that time) were joined together at their ends and separated along their length, a magnetic field developed around the two legs, provided however that a temperature difference existed between the two junctions. He published his observations the following year, and the phenomenon came to be known as the Seebeck effect. The significance of his discovery notwithstanding, Seebeck did not correctly identify the cause of the magnetic field. The magnetic field results from an equal but opposite electric current in the leg of each metal strip caused by a thermally generated electric potential difference between the junctions. If one junction is broken but the temperature differential is maintained, current no longer flows in the legs but a voltage can be measured. This generated voltage, V, is the Seebeck voltage and is related to the difference in temperature, ΔT, between the heated junction and opened junction by a proportionality factor, α, called the Seebeck coefficient, or $V = \alpha\Delta T$. The value for α is dependent on the types of material at the junction.

Seebeck effect

In 1834 the French physicist and watchmaker Jean-Charles-Athanase Peltier observed that if a current is passed through a single junction of the type described above, the amount of measured heat generated is not consistent with that which would be predicted from Joule heating (see below) alone. This observation is called the Peltier effect. As in Seebeck's case, Peltier failed to define the cause of the anomaly. He did not identify that heat was absorbed or evolved at the junction depending on the direction of current. He also did not recognize the reversible nature of this thermoelectric phenomenon and associate his discovery with Seebeck's.

Peltier effect

It was not until 1855 that William Thomson (later Lord Kelvin) drew the connection between the Seebeck and Peltier effects and made a significant contribution to the understanding of thermoelectric phenomena. The Peltier heat, Q_p, was shown to be proportional to the applied junction current, I, through the relationship $Q_p = \pi I$, where π is the Peltier coefficient. Thomson showed through thermodynamic analysis that $\pi = \alpha T$, where T is the absolute temperature of the junction. The Thomson effect, theoretically predicted by Thomson on the basis of thermodynamic considerations, showed that heat is absorbed or evolved, Q_τ, along the length of a material rod whose ends are at different temperatures. Q_τ was shown to be proportional to the flow of current, I, and the temperature gradient along the rod. The proportionality factor, τ, is known as the Thomson coefficient.

Thomson effect

All thermoelectric phenomena are described by these three effects. Analysis of a thermoelectric device is, however, adequately performed using only one of the thermoelectric parameters, the Seebeck coefficient, α. The reason is that the Thomson effect is small, and so it is generally neglected. The Peltier coefficient, on the other hand, is related to α through the operating condition of the junction temperature.

Two nonthermoelectric quantities must also be identified before a thermoelectric device can be appropriately described. They are Joule heating (the production of heat in a conductor when a current flows through it, as in the case of filaments of an electric kitchen range or toaster) and thermal conduction (the transfer of heat due to tem-

perature differences between adjacent parts of a body). Although a thermoelectric device is made up of many *p*-type and *n*-type semiconductor legs, its behaviour can be discussed using only one couple.

Figure 79 shows a *p*-type and *n*-type semiconductor leg coupled to a heat source, heat sink, and an electrical power consuming load. (Other couples can be connected electrically in series and thermally in parallel.) The leg geometry affects operation. The leg length is L, and the base area, a, is w^2. Under the condition that *p*- and *n*-type semiconductors are similar in their measured properties, average value parameters can be used to analytically describe the couple. The heat flow through the couple at T_1 is given by

$$H = 2aIT - I^2\rho\left(\frac{L}{a}\right) + 2\kappa\left(\frac{a}{L}\right)\Delta T,$$

where temperature is in kelvins, ρ is the electrical resistivity in ohms-centimetre, κ is the thermal conductivity in watts per centimetre kelvin, a is microvolts per kelvin, and L/a is in centimetres^{-1}. In this equation, the first term results from the reversible Peltier effect that generates heat at the top junction. The second term reflects loss due to irreversible Joule heating (one half of the total amount generated). The last term is the irreversible heat loss due to thermal conductivity in each leg.

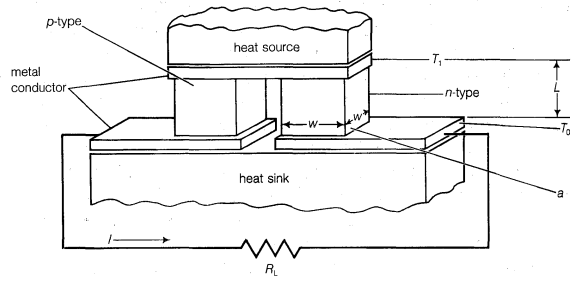


Figure 79: Single couple of a thermoelectric generator.

In a thermoelectric power generator, a temperature differential between the upper and lower surfaces of two legs of the device results in power being generated. If a power consuming load is not attached to the generator (open-circuited), the applied heat source (H) results in a temperature differential (ΔT) of some value dictated only by the thermal conductivity of the *p*- and *n*-type semiconductor legs. Since no current would flow in the thermoelectric device, no power would be generated. (The first and second terms of the above equation would be zero.) Because of the Seebeck effect, however, a voltage would be present at the output terminals, just like in an unconnected battery. When a load is attached, current will flow through the load. The Seebeck voltage $V_a = a\Delta T$ is divided between two terms: the internal device voltage drop IR_{int} due to internal resistance $R_{int} = 2\rho(L/a)$ (for the couple), and the external voltage drop IR_L. It is the Seebeck voltage and these two resistances that dictate the flow of current (and the generated output electrical power) given by

$$I = \frac{2a\Delta T}{(R_{int} + R_L)}.$$

This same current pumps heat within the thermoelectric device due to the Peltier effect, which in turn results in a lowering of the initial temperature differential when the current is zero. Part of the heat energy, H, through the Seebeck generated current, is converted to Joule heating within the legs of the thermoelectric device. The efficiency, η, for a power generator is the output power, I^2R_L, divided by H. It can be shown that

Thermal efficiency

$$\varepsilon \text{ max} = \left(\frac{T_1 - T_0}{T_1}\right)\left(\frac{\sqrt{1 + Z\overline{T}} - 1}{\sqrt{1 + Z\overline{T}} + \frac{T_0}{T_1}}\right),$$

where the first term is the Carnot efficiency (see *Transformation of energy* above). The second term contains $\overline{T}$, which is the average temperature of the leg. The Z is the figure of merit of the semiconductor legs; it represents a "quality factor" of the material to perform as thermoelectric device (it is 3×10^{-3} per kelvin at 300 K), given by

$$Z = \frac{a^2}{\kappa\rho}.$$

For material quality to improve (*i.e.*, larger Z), it is generally agreed that the thermal conductivity (κ) and electrical resistivity (ρ) of semiconductor materials must decrease. This has been the principal limiting factor toward higher conversion efficiency in thermoelectric power generation, which in turn has limited the use of thermoelectric devices. A new effort in materials research is required to obtain materials that can improve the overall efficiency of thermoelectric devices.

Major types of thermoelectric generators. Thermoelectric power generators vary in geometry, depending on the type of heat source and heat sink, power requirement, and intended use. In general, many units require a power conditioner to convert the generator output to a usable voltage value. Although the Soviet army used these devices to power portable communications transmitters during World War II, modern power generators are based on the substantial improvements made in semiconductor materials and electrical contacts between 1955 and 1965, as well as on engineering improvements achieved up to the present.

Fossil-fuel generators. Units have been constructed to use natural gas, propane, butane, kerosene, jet fuels, and wood, to name but a few heat sources. A 500-watt multifuel, maintenance-free tactical power generator for advance area application has been developed for the U.S. Army. Commercial units are in the 10- to 100-watt output power range for use in remote areas. Applications for these units include navigational aids, data collection systems and communications systems, and cathodic protection, which prevents electrolysis from corroding metallic pipelines and marine structures.

Solar-source generators. Early attempts to construct solar thermoelectric generators for orbiting spacecraft failed because of low efficiency and higher unit weight compared to silicon solar cells. They have, however, been used with some success to power small irrigation pumps in remote areas and underdeveloped regions of the world where fuel sources are unreliable. In addition, a group of U.S. researchers have described an experimental system capable of using warm surface ocean water as the heat source and cooler deep ocean water as the heat sink for large power generation. Economics favouring this system are based on it being so reliable that there is minimal maintenance cost. Still another system design features both heat pumping and power generation for thermal control of orbiting spacecraft. Utilizing solar heat from the Sun-oriented side of the spacecraft, thermoelectric devices generate electrical power. This power is used to supply current to other thermoelectric devices in dark areas of the spacecraft to reject heat from the vehicle. Operating in this mode, the use of thermoelectric devices, with their reversible function capability, decreases the amount of power required by the spacecraft to increase overall heat expulsion.

Nuclear-fueled generators. Thermoelectric generators that use radioisotopes as fuel derive a high-temperature heat source by the self-absorption of emitted decay products. Because thermoelectric devices are relatively immune to nuclear radiation and because the source can be made to last for a long period of time, such generators provide a unique source of power for many unattended and remote applications. For example, radioisotope thermoelectric generators provide electric power for nonorbiting as well as Earth-orbiting spacecraft, instrumentation for deep-ocean data collection and surface monitoring, warning and communications systems, isolated terrestrial weather monitoring stations, and certain medical applications. A low-power radioisotope thermoelectric generator was developed as early as 1970 and used to power cardiac pacemakers. The power range of radioisotope thermoelectric generators is between 10^{-6} and 10^2 watts.

Applications of radioisotope thermoelectric generators

Development of thermoelectric power generators. The

first application of a thermoelectric generator was in all likelihood Peltier's use of the Seebeck effect (see above) to generate a small amount of power required to pump heat in his junction experiments. An understanding of the principle involved in this phenomenon led to the use of dissimilar metal wires for measuring temperature—namely, the thermocouple. From this evolved the use of multiple but alternating dissimilar metallic wires in a thermopile with which to measure optical radiation.

As the need for electric power became increasingly more important between 1885 and 1910, investigators began studying thermoelectricity systematically. By 1910 E. Altenkirch, a German scientist, satisfactorily calculated the efficiency of thermoelectric generators and delineated the parameters of the materials needed to build practical devices. Unfortunately metallic conductors were the only materials available at the time, rendering it unfeasible to build thermoelectric generators with an efficiency of more than 0.6 percent.

During the late 1920s, Soviet researchers actively pursued theoretical and experimental work on thermoelectricity because of the need for electric power in remote yet habitable areas of their vast country. By 1940 a unit with a conversion efficiency of 4 percent had been developed using semiconductors. It was quickly realized that semiconductor materials were best suited for thermoelectric application. By the early 1950s, interest in thermoelectric power generation was on the rise in certain highly industrialized nations, most notably the United States. Scientific projects being undertaken by these countries in isolated, uninhabited areas necessitated power sources for data collection and communications systems. Yet, in spite of the increased research and developmental activity, gains in thermoelectric power-generating efficiency were relatively small. An efficiency capability of not much more than 10 percent had been attained as of the late 1980s. Better thermoelectric materials are required to go much beyond this performance level. Still, some varieties of thermoelectric generators have proved to be of considerable practical import. Those fueled by radioisotopes are the most versatile, reliable, and generally used power source for isolated or remote sites. (J.W.H.)

Adoption of semiconductors

THERMIONIC POWER CONVERTERS

General characteristics. A thermionic power converter—also variously called thermionic generator, thermionic power generator, or thermoelectric engine—is a device in which heat energy is directly converted into electrical energy. It has two electrodes. One of these is raised to a sufficiently high temperature to become a thermionic electron emitter and can be dubbed the "hot plate." The other electrode, called a collector because it receives the emitted electrons, is operated at a significantly lower temperature. The space between the electrodes is normally filled with a vapour or gas at low pressure (on the order of 1.333×10^2 pascals). The thermal energy may be supplied by chemical, solar, or nuclear sources.

Primary components

Principles of operation. The emission of electrons from the hot plate is analogous to the liberation of steam particles when water is heated. The flow of electrons may be completed by interconnecting the two electrodes by an external load, shown by a resistor R_L in Figure 80. Part of the thermal energy that is supplied to liberate the electrons ("boil them off") is converted directly into electrical energy.

A thermionic power converter can be viewed in several different ways. It can, for example, be examined in terms of thermodynamics as a heat engine that utilizes an electron-rich gas as its working fluid. A thermionic converter also may be thought of as a thermoelectric device—a thermocouple in which one of the conductors has been replaced by either a plasma or a vacuum (*i.e.*, an evacuated space). It can even be regarded in terms of electronics as a diode that converts heat to electrical energy via thermionic emission. No matter how thermionic converters are conceived of or labeled, however, they all work due to the discharge of electrons from heated conducting materials. The following discussion treats devices of this sort as heat engines.

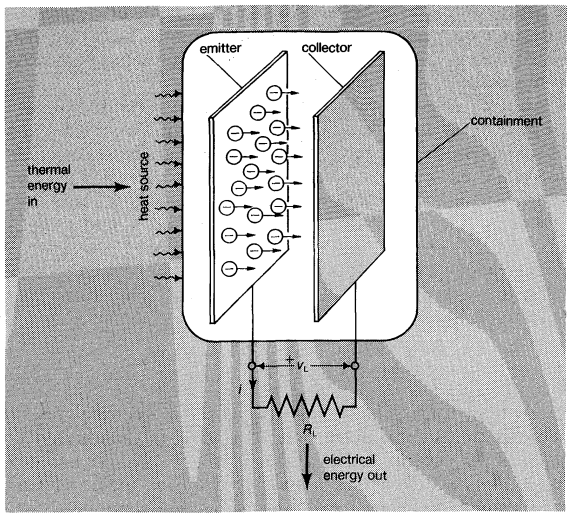


Figure 80: Schematic of a basic thermionic converter.

From E.M. Walsh, *Energy Conversion*, copyright © 1967 by the Ronald Press Company; reprinted by permission of John Wiley & Sons, Inc.

The major problem in developing large-scale thermionic power converters is the limit imposed on maximum current density due to the space-charge effect—*i.e.*, the negatively charged electrons that are emitted deter the movement of other electrons toward the collecting electrode. Two solutions to this problem have been pursued. One involves reducing the spacing between the electrodes to the order of micrometres, while the other entails the introduction of positive ions into the cloud of negatively charged electrons in front of the emitter. The latter method has proved to be the most feasible from many standpoints, especially manufacturing. It has resulted in the development of both the cesium and the auxiliary discharge thermionic power converters.

Minimizing space-charge effect

Thermionic emission. The emission of electrons is fundamental to thermionic power conversion. The mechanism for the escape of an electron is shown in Figure 81. The actual effect of a negatively charged electron (Figure 81A) may be represented equivalently by a positively charged electron located in a mirror-image arrangement (Figure 81B). This model permits the escape force to be determined from a fundamental law of physics, the inverse square law. That force is given by

$$\frac{e^2}{16\pi\varepsilon_0 x^2},$$

where e is electronic charge (coulombs) and ε_0 is permittivity of free space. The energy required to overcome this force—to cause the electron to escape—is called the work function φ. Each material has a unique value, as shown in Table 4, at common emitter temperatures above 2,000 K. (Collectors normally operate around 1,000 K.) The other parameter tabulated, R, is material-dependent, although the theoretical derivation of the governing equation fixes

From S.W. Angrist, *Direct Energy Conversion*, copyright © 1982 by Allyn and Bacon, Inc.

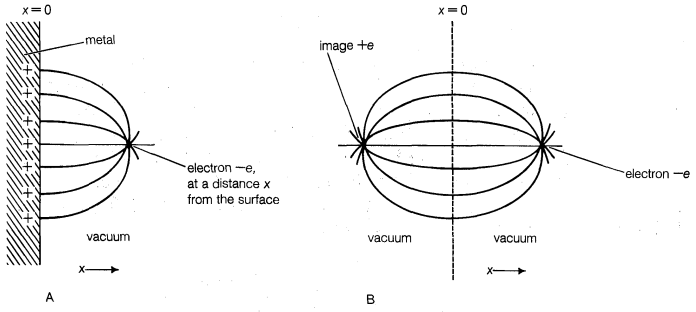


Figure 81: *Mechanism for electron escape in thermionic power conversion.*
(A) The electric field lines for an electron near the surface of a metal. (B) Electric field lines for an image charge $+e$ and an electron at equal distances on either side of $x = 0$. The field for x greater than zero is identical with the field A (see text).

its value as a universal constant $R = 1.2 \times 10^{-6}$ amperes per square metre kelvin squared (amp/m²-K²).

The rate at which electrons are liberated from the surface of the emitter is given by the Richardson–Dushman electron current density equation; *i.e.*,

Richardson–Dushman equation

$$J_0 = RT^2 \exp\left(-\frac{e\varphi}{kT}\right),$$

where T is absolute temperature (K) and k is Boltzmann's gas constant for one molecule (ergs per kelvin). This equation for emission current is named for Owen Willans Richardson and Saul Dushman, who did pioneering work on the phenomenon. The rate of emission increases rapidly with temperature and decreases exponentially with the work function. It is always desirable to operate a thermionic converter at a high temperature as well as to be selective in choosing its electrode material.

When electrons escape the emitter surface, they gain energy equal to the work function with some excess kinetic energy. Upon striking the collector, their kinetic energy is used to "absorb" the electrons into the surface. This absorbed energy must be rejected as heat from the collector or force the electrons through the external load, thereby giving the desired electrical energy conversion.

Table 4: Thermionic Emission Properties of Certain Materials

material	φ (volts)	R (amp/m²-K²) $\times 10^{-6}$
Cesium	1.89	0.5
Molybdenum	4.2	0.55
Nickel	4.61	0.3
Platinum	5.32	0.32
Tungsten	4.52	0.6
Tungsten + cesium	1.5	0.03
Tungsten + barium	1.6	0.015
Tungsten + thorium	2.7	0.04
Barium oxide	1.5	0.001
Strontium oxide	2.2	1.0

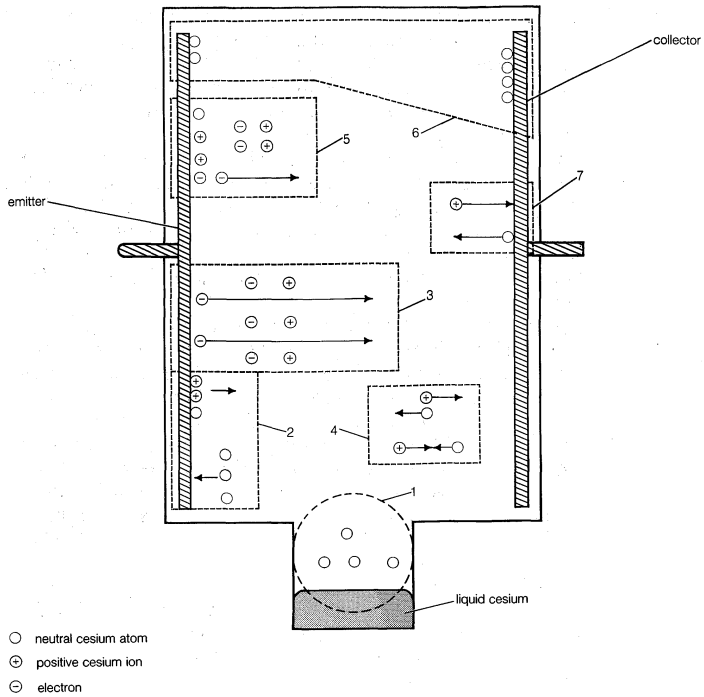


Figure 82: *The various processes in a gas-filled converter.* They occur in the following sequence: (1) evaporation of liquid cesium, (2) arrival of cesium atoms at emitter and departure as ions, (3) neutralization of space charge, (4) energy sharing of ions with atoms, (5) formation of an ion space-charge sheath, (6) reduction of work function due to cesium deposition, and (7) cesium ion recombination at the collector surface.

From S.W. Angrist, *Direct Energy Conversion*, copyright © 1982 by Allyn and Bacon, Inc.

Major types of thermionic converters. *Vacuum converters.* This type of thermionic device has a vacuum gap between its electrodes. Because of the small spacing required between the emitter and collector to counteract the space charge, the vacuum converter has had only limited practical application; however, it has given rise to other configurations of greater utility. They are briefly described below.

Gas-filled converters. These devices are designed in such a way that positively charged ions are continuously generated and mixed with negatively charged electrons in front of the emitter to neutralize the electrostatic field. Because of this, a liberated electron has no electrostatic resistance in passing from the emitter to the collector. Figure 82 shows schematically the operation of a cesium-filled converter. Cesium is used in the most efficient converters because of its low ionization potential (3.87 electron volts). Potassium, rubidium, and various other metals produce similar results. The arrival rate of neutral cesium atoms is dependent on the gas pressure of cesium and its reservoir temperature. For efficient production of ions, the emitter temperature should be approximately 3.6 times the reservoir temperature.

Auxiliary discharge converters. Such thermionic generators operate at lower temperatures (say, 1,500 K), permitting the use of a fossil-fuel heat source. Ions are produced by applying voltage to a third electrode, shown schematically as auxiliary anodes in Figure 83. The gas between the electrodes in this system is inert (e.g., neon, argon, or xenon). The principal advantage of the auxiliary discharge converter—so called because of its spark plug-type configuration—is that conventional fossil fuels are adequate for the heat source. The disadvantage is the complexity of the discharge system.

Fossil fuels as heat source

Because thermionic converters are tolerant of high accelerations, have no moving parts, and exhibit a relatively high power-to-weight ratio, they are well suited for applications in spacecraft. Since they function best at high temperatures, they may be used as topping devices (*i.e.*, power boosters) on conventional power plants. Their efficiencies make them suitable power sources for remote or hostile environments (*e.g.*, under water) or for use in low-power radio transmitters.

Development of thermionic devices. As early as the mid-18th century, Charles François de Cisternay Du Fay, a French chemist, noted that electricity may be conducted in the gaseous matter—that is to say, plasma—adjacent to a red-hot body. In 1853 the French physicist Alexandre-Edmond Becquerel reported that only a few volts were required to drive electric current through air between high-temperature platinum electrodes. From 1882 to 1889 Julius Elster and Hans Geitel of Germany perfected a sealed device containing two electrodes, one of which could be heated while the other one was cooled. They discovered that, at fairly low temperatures, electric current flows with little resistance if the hot electrode is positively charged. At moderately higher temperatures, current flows readily in either direction. At even higher temperatures, however, electric charges from the negative electrode flow with the greatest ease.

In the 1880s the American inventor Thomas A. Edison applied for a patent pertaining to thermionic emission in a vacuum. In his patent request, he explained that a current

From E.M. Walsh, *Energy Conversion*, copyright © 1967 by The Ronald Press Company; reprinted by permission of John Wiley & Sons, Inc.

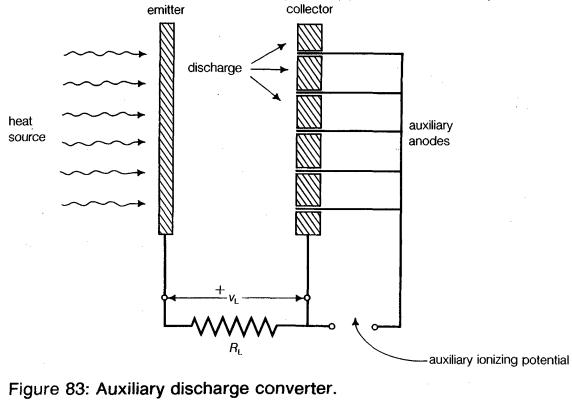


Figure 83: Auxiliary discharge converter.

Edison effect

passes from a heated filament of an incandescent electric lamp to a conductor in the same glass globe. Though Edison was the first to disclose this phenomenon, which later came to be known as the Edison effect, he made no attempt to exploit it; his interest in perfecting the electric light system took precedence.

In 1899 the English physicist J.J. Thomson defined the nature of the negative charge carriers. He discovered that their ratio of charge to mass corresponded to the value he found for electrons, giving rise to an understanding of the fundamentals of thermionic emission. In 1915 W. Schlichter proposed that the phenomenon be used for generating electricity.

By the early 1930s the American chemist Irving Langmuir had developed sufficient understanding of thermionic emission to build basic devices, but little progress was made until 1956. That year another American scientist, George N. Hatsopoulos, described in detail two kinds of thermionic devices. His work led to rapid advances in thermionic power conversion. Recent research has been centred primarily on a converter capable of utilizing thermal energy from a nuclear reactor on board spacecraft.

(L.E.Si.)

MAGNETOHYDRODYNAMIC POWER GENERATORS

General characteristics. Magnetohydrodynamic (MHD) power generators produce electrical power through the interaction of a flowing, electrically conducting gas (or other fluid) and a magnetic field. Various countries, including the Soviet Union, Japan, China, Poland, and the United States, have undertaken active developmental programs, since MHD power plants offer the potential for large-scale electrical power generation at reasonable cost with comparatively little detrimental impact on the environment. Generators of the MHD type are also attractive for the production of large electrical power pulses, and their first practical application has been for this kind of service (see below).

The underlying principle of MHD power generation is elegantly simple. An electrically conducting fluid is driven by a primary energy source (*e.g.*, combustion of coal or a gas) through a magnetic field, resulting in the establishment of an electromotive force within the conductor in accordance with the principle established by Faraday (see above). Furthermore, if the conductor is an electrically conducting gas, it will expand, and so the MHD system constitutes a heat engine involving an expansion from high to low pressure in a manner similar to that of a gas turbine. The MHD system, however, involves a volume interaction between a gas and the magnetic field through which it is passing (see below), whereas the gas turbine operates through the gas interaction with the surfaces of a rotating blade system. It is, in effect, a system that depends on volume rather than surface interaction.

The MHD generator as an electromagnetic turbine

The MHD generator can properly be viewed as an electromagnetic turbine because its output is obtained from the conducting gas–magnetic field interaction directly in electrical form rather than in mechanical form, as in the case of a gas (or steam) turbine. This is illustrated in Figure 84, which compares a conventional turbogenerator with an MHD system. Other types of MHD turbines are possible and will be mentioned below. Here, attention is concentrated on the electrically conducting gas type, which has been the focus of most research and developmental work.

Electrical conduction in gases occurs when electrons are available to be organized into an electric current in response to an applied or induced electric field. The electrons may be either injected or generated internally, and, because of the electrostatic forces involved, they require the presence of corresponding positive charge from ions to maintain electrical neutrality. An electrically conducting gas consists in general of electrons, ions to balance the electric charge, and neutral atoms or molecules. Such a gas is termed a plasma.

In MHD generators, electrons for supporting the flow of current can be obtained in either of two ways: by heating the gas to a sufficiently high temperature to yield electrons through ionization or by the induction of a sufficiently strong electric field in a manner similar to that in gas-discharge devices. These methods are referred to as thermal ionization and nonequilibrium ionization, respectively. In either case, the mechanism of energy transfer from the flowing fluid to the electrical output can be thought of as a coupling of the electron-comprised gas to the ions through electromagnetic forces; the ions in turn are embedded in the background of atomic or molecular gas and lack mobility by virtue of their being coupled to the molecules or ions through collision processes described by kinetic behaviour.

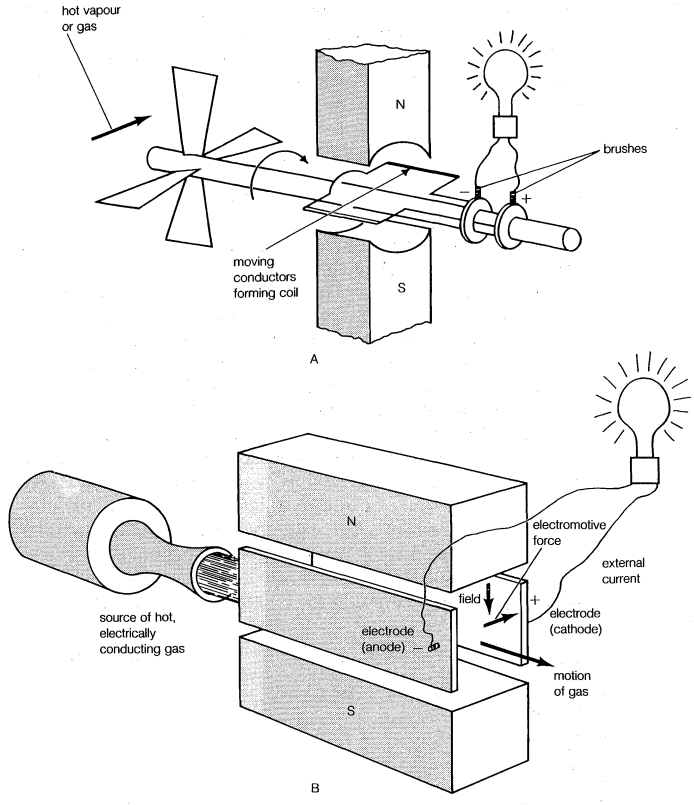


Figure 84: *Comparison of the operating principles of a turbogenerator and an MHD generator.*
(A) Turbogenerator and (B) MHD generator.

Interest in MHD-power generation was originally stimulated by the observation that the interaction of a plasma with a magnetic field could occur at much higher temperatures than were possible in a system consisting of a rotating mechanical system. The limiting performance from the point of view of efficiency in heat engines is established by the Carnot efficiency, obtained from the difference between the absolute hot source temperature, T_1, and the cold sink temperature, T_0, divided by T_1. For example, when the source temperature is 2,810 K and the sink temperature is that of the environment (say, 294 K), the Carnot efficiency is slightly less than 90 percent. Allowing for the inefficiencies introduced by finite heat transfer rates and component inefficiencies in real heat engines, a system employing an MHD generator offers the potential of an ultimate efficiency in the range 60 to 65 percent. This is to be compared with 35 to 36 percent achieved by a modern coal-fired, steam-turbine plant with scrubbers (devices that absorb sulfur dioxide from exhaust gases); 40 percent with a natural gas-fired, steam-turbine plant; and about 46 percent projected for gas-fired, combined gas–steam turbine installations. The implications of this efficiency improvement are an enhanced utilization of primary fuel resources due to higher thermodynamic efficiency and a lower emission of environmental pollutants. (The environmental advantages are discussed in *Major types of MHD systems* below.)

Comparatively high conversion efficiency of MHD generators

Principles of operation. As in the case of all electrical machines, the power output of MHD generators for every cubic metre of conductor depends directly on its conductivity, the square of the velocity at which the conductor

moves, and the square of the magnetic field through which it is passing. For MHD generators to operate competitively, the electrical conductivity of the plasma must be adequate to achieve good performance and reasonable physical dimensions in the temperature range of about 1,800 K and upward—*i.e.,* temperatures at which the turbine blades of a gas-turbine power system would no longer be able to operate. Analysis shows, and experience confirms, that adequate conductivity results if a small amount of additive, typically around 1 percent by mass, is injected into the working gas of the MHD system. This additive is in the form of readily ionizable material such as potassium carbonate and is referred to as the "seed." It is the principal source of electrons (and ions) that render the gas electrically conducting and thereby enable direct conversion to occur.

Addition of a seed material

The hot gas, at a pressure of several megapascals, has seed material added and is accelerated by a nozzle to a speed usually greater than that of sound (*i.e.,* to supersonic conditions). As shown in Figure 85, it then enters a containment structure known as the channel, or duct, across which a powerful magnetic field is applied. In accordance with the Faraday induction principle, an electromotive force acting in a direction perpendicular to the flow and field is set up and, to enable this to provide a current to an external circuit, the walls parallel to the magnetic field serve as electrodes. Because the electromagnetic force is induced in the gas, the positive electrode is the cathode, or electron emitter, and the negative electrode is the anode, or collector (Figure 85). The remaining two walls of the channel are insulators that confine the resultant voltage. Depending on the heat source and magnetic field strength, power densities of 10 to 500 megawatts per cubic centimetre in the duct can be obtained. A magnetic field in the range 4.5 to 6 teslas is required to achieve these values, and this is most readily obtained by using a superconducting magnet.

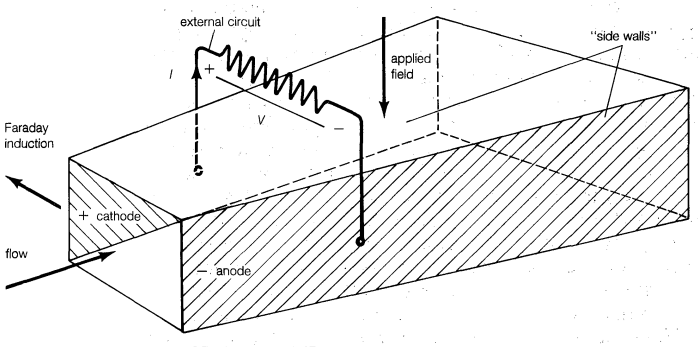


Figure 85: *Simple MHD generator.*
The load current is represented by *I* and the voltage by *V* (see text).

A complicating feature of a plasma MHD generator is the occurrence of a pronounced Hall effect, which results from the behaviour of electrons in the presence of both magnetic and electric fields. Electrons are accelerated in the direction of an electric field but follow a circular path around a magnetic field line (cyclotron behaviour). When these two actions are combined and the collision processes taken into account, the effect (named after its discoverer, the American physicist Edwin H. Hall) is for the electric current to flow at an angle with respect to the electric field, producing an additional field along the axis of the MHD duct. This field, called the Hall field, causes an axial current (Hall current) to flow if the electrodes are continuous, as in Figure 85. This in turn requires that either the electrode walls be constructed to support the Hall field or that the Hall field itself be used as the output to drive current through the electric circuit external to the MHD system.

A number of generator configurations can be used to achieve this objective. The principal ones are briefly described here. In the so-called Faraday generator (Figure 86A), the electrode walls are segmented to support the axial potential, and the power is taken out in a series of loads. The Hall generator (Figure 86B) maximizes the Hall output by short-circuiting the Faraday terminals and connecting a simple load between the ends of the duct. Consideration of the potentials at different points in the duct have led to the conclusion that an equipotential runs diagonally (across the insulator walls) and that, accordingly, electrodes may be connected along such a potential to achieve the diagonal configuration shown in Figure 86C. This diagonal generator may be thought of as a Faraday type in which the individual electrode pairs have been connected in series in a manner that does not violate the potential required for correct operation of the duct yet permits a single load to be used.

MHD generator configurations

An attractive alternative to the linear Hall generator in Figure 86B is the disk generator in which a radial output flow occurs and the short-circuited Faraday currents flow in closed circular paths (Figure 86D). The Hall output appears between the centre and the periphery of the disk. This disk generator is particularly attractive when nonequilibrium ionization is employed.

Major types of MHD systems. The type of ionization employed by an MHD power generator depends on the heat source selected and the method used to couple it to the working fluid. Several possibilities exist. A complete MHD system may include a solid- or liquid-fuel rocket motor (see *Rockets* above), seed injector, nozzle, duct, and magnet and may utilize thermal ionization in the combustion products that make up the working fluid. MHD generators currently in service are of this type. They are compact systems capable of providing very large amounts of power. Natural gas, oil, and coal also are excellent potential fuels for MHD systems and were in fact the first to be proposed and considered. With the addition of oxygen or compressed preheated air or both, these fossil fuels yield combustion products that readily reach the temperatures required for thermal ionization.

Possible fuel sources

Although conventional nuclear fission reactors of the light-water type operate at temperatures too low for MHD applications, nuclear heat sources represent still another option for MHD systems. If a nuclear heat source were employed, hydrogen or a noble gas such as argon or helium would be appropriate for the working fluid, and nonequilibrium ionization could be used. A possible candidate for this kind of heat source is the NERVA (nuclear energy for rocket vehicle application) high-temperature fission reactor, originally designed for space propulsion. While the ultimate form of fusion reactor has yet to be determined, it should be feasible to devise a scheme for coupling an MHD generator to a nuclear source of this type (see below). Solar concentrators also can in theory achieve the temperatures required for MHD operation, and there have been several proposals for exploiting solar radiation to provide the necessary thermal energy.

The use of fission and fusion reactors as heat sources for MHD generators is contingent upon the development of suitable high-temperature reactor systems. Similarly, in the case of solar-based MHD, high-temperature collectors for solar thermal systems are required. Since such systems have yet to be constructed, attention has so far been focused on fossil- and chemical-fueled systems, with the primary aim of using MHD technology for central station power generation.

As energy is extracted from an MHD generator, duct conditions become increasingly less favourable for maintenance of electrical conductivity and, in the case of thermal ionization, extraction is essentially completed when the temperature falls to about 2,500 K. A central station power system thus has to be based on a binary cycle, with an MHD generator topping a conventional steam plant. (Topping means that the gas generated by burning a fossil fuel is first passed through the MHD generator and then on to the turbogenerator of the ordinary power plant, which constitutes the bottoming portion of the binary cycle.) In effect, the exhaust gas from the MHD generator feeds the bottoming cycle so that the residual thermal energy in the gas can be used to furnish additional power output and also to preheat the oxidizer for further fuel combustion in the MHD generator. An MHD power plant employing such an arrangement is an open-cycle ("once-through") system.

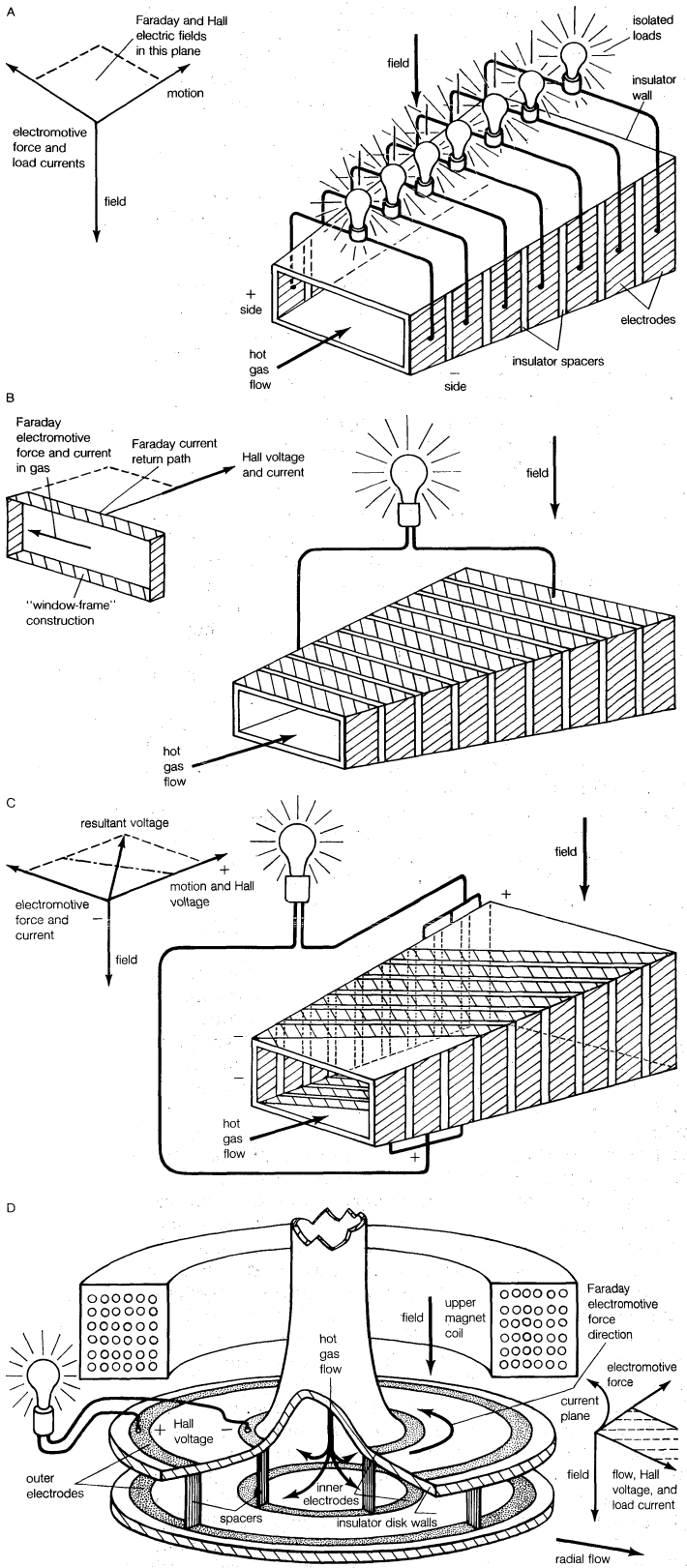


Figure 86: *MHD generator configurations.*
(A) Segmented Faraday generator, (B) Hall generator, (C) diagonal generator with "window-frame" construction, and (D) disk generator (see text).

The abundance of coal reserves in the United States has favoured the development of coal-fired MHD systems for domestic use. Coal combustion as a source of heat has several advantages. For example, it results in coal slag, which under magnetohydrodynamic conditions is molten and provides a layer that covers all of the insulator and electrode walls. The electrical conductivity of this layer is sufficient to provide conduction between the working gas and the electrode structure but not so high as to cause any significant leakage of electric currents and consequent loss. Indeed, the reduced thermal loss to the walls due to the slag layer more than compensates for any electrical losses arising from its presence.

Environmental benefits of coal-fired MHD power plants

The use of a seed material in conjunction with coal firing offers environmental benefits. When sulfur is present in the coal, the recombination chemistry that occurs in the duct of an MHD generator as the gas proceeds to lower temperatures favours the formation of potassium sulfate and so facilitates the removal of sulfur from high-sulfur coals. This in turn sharply reduces sulfur dioxide emissions. Moreover, the need to recover seed material ensures that a high level of particulate removal is built into an MHD coal-fired plant. Finally, by careful design of the boiler and control of combustion, very low levels of nitrogen oxides can be achieved. From an environmental viewpoint, MHD power systems offer effluent levels impressively lower than those currently established by the Environmental Protection Agency (EPA) in the United States.

Use in geophysical research

The need to provide large pulses of electrical power at remote sites has stimulated the development of pulsed MHD generators. For this application, the MHD system consists basically of a rocket motor, duct, magnet, and connections to electrical load. Similar generators are routinely operated in the Soviet Union as sources for pulse-power electromagnetic sounding apparatuses used in geophysical research; power levels up to 100 megawatts have been reported. In this application, the MHD generator provides a power pulse typically of a few seconds' duration to a magnetic or electric dipole located on the surface. The magnetic fields induced in the crust of the Earth are measured, and, through electrical conductivity, properties of the crust are determined.

An alternative MHD scheme involves a generator of the type shown in Figure 85, except that it employs a liquid metal as its electrically conducting medium. Liquid metal is an attractive option because of its high electrical conductivity, but it cannot serve directly as a thermodynamic working fluid. The liquid has to be combined with a driving gas or vapour to create a two-phase flow in the generator duct, or it has to be accelerated by a thermodynamic pump (often described as an ejector) and then separated from the driving gas or vapour before it passes through the duct. Depending on whether a condensable vapour or a gas is used, a number of cycles is possible, including condensing cycles essentially similar to that employed in a steam turbine. While the so-called liquid metal MHD systems offer attractive features from the viewpoint of electrical machine operation, they are limited in temperature by the properties of liquid metals to about 1,250 K. They thus have to compete with various existing energy-conversion systems and with other advanced systems capable of operating in the same temperature range.

The use of MHD generators to provide power for spacecraft for both burst and continuous operations has been considered. While both chemical and nuclear heat sources have been investigated, the latter is the preferred choice for applications such as supplying electric propulsion power for deep-space probes.

Development of MHD power generators. The first recorded MHD investigation was conducted in 1821 by the English chemist Humphry Davy when he showed that an arc could be deflected by a magnetic field. More than a decade later, Faraday sought to demonstrate motional electromagnetic induction in a conductor moving through the magnetic field of the Earth. To this end he set up in January 1832 a rudimentary open-circuit MHD generator, or flow meter, on the Waterloo Bridge across the River Thames. His experiment was unsuccessful, however, due to the electrodes being electrochemically polarized, an effect not understood at that time.

Faraday soon turned his attention to other aspects of electromagnetic induction, and MHD power generation received little attention until the 1920s and '30s, at which time B. Karlovitz, a Hungarian-born engineer, first pro-

posed a gaseous MHD system of the type described above. In 1938 he and D. Halász set up an experimental MHD facility at the Westinghouse research laboratories and by 1946 had shown that, through seeding the working gas, small amounts of electric power could be extracted. The project was abandoned, however, largely because of a lack of understanding of the conditions required to make the working gas an effective conductor.

First successful MHD power generator

Interest in magnetohydrodynamics grew rapidly during the late 1950s as a result of extensive studies of ionized gases for a number of applications. In 1959 the American engineer Richard J. Rosa operated the first truly successful MHD generator; this device produced about 10 kilowatts of electric power. By 1963 the Avco Research Laboratory, under the direction of the American physicist Arthur R. Kantrowitz, had constructed and operated a 33-megawatt MHD generator, and for many years this remained a record power output. The assumption in the late 1960s that nuclear power would dominate commercial power generation and the failure to find applications for space missions led to a sharp curtailment of MHD research. The energy crisis of the 1970s, however, brought about a revival, with the focus centred on coal-fueled systems in the United States and various other countries. By the late 1980s, development had reached the point where the construction of a complete demonstration system was feasible and, with the environmental advantages resulting from efficient conversion becoming increasingly apparent, the incentive to construct such a system within the next decade gained impetus. (W.D.J.)

FUSION REACTORS

Since the 1930s, scientists have known that the Sun and other stars generate their energy by nuclear fusion. They realized that if fusion energy generation could be replicated in a controlled manner on Earth, it might very well provide a safe, clean, and inexhaustible source of energy. The 1950s saw the beginning of a worldwide research effort to develop a fusion reactor. The substantial accomplishments and prospects of this continuing endeavour are described here.

General characteristics. The energy-producing mechanism in a fusion reactor is the joining together of two light atomic nuclei. When two nuclei fuse, a small amount of mass, m, is converted into a large amount of energy, E. Energy and mass are related through Einstein's relation, $E = mc^2$, by the large conversion factor c^2, where c is the speed of light. The inverse process, conversion of mass to energy by the splitting of a heavy nucleus, is the basis for the fission reactor (see *Nuclear fission reactors* above).

Energy-producing mechanism in fusion reactors

Fusion reactions are inhibited by the electrical repulsive force that acts between two positively charged nuclei. For fusion to occur, the two nuclei must approach each other at high speed to overcome the electrical repulsion and attain a sufficiently small separation (less than one-trillionth of a centimetre) that the short-range strong nuclear force dominates. For the production of useful amounts of energy, a large number of nuclei must undergo fusion; that is to say, a gas of fusing nuclei must be produced. In a gas at extremely high temperature, the average nucleus contains sufficient kinetic energy to undergo fusion. Such a medium can be produced by heating an ordinary gas of neutral atoms beyond the temperature at which electrons are knocked out of the atoms. The result is an ionized gas consisting of free negative electrons and positive nuclei. This gas constitutes a plasma. Most of the matter in the universe is in the plasma state.

The scientific problem of fusion is thus the problem of producing and confining a hot, dense plasma. The core of a fusion reactor would consist of burning plasma. Fusion would occur between the nuclei, with the electrons present only to maintain macroscopic charge neutrality.

Stars, including the Sun, consist of plasmas that generate energy by fusion reactions. In these "natural fusion reactors" the reacting, or burning, plasma is confined by its own gravity. It is not possible to assemble on Earth a plasma sufficiently massive to be gravitationally confined. The hydrogen bomb is an example of fusion reactions produced in an uncontrolled, unconfined manner in which the energy density is so high that the energy release is explosive. By contrast, the use of fusion for peaceful energy generation requires control and confinement of a plasma at high temperature and is often called controlled thermonuclear fusion.

In the development of fusion power technology, demonstration of "energy breakeven" is taken to signify the scientific feasibility of fusion. At breakeven, the fusion power produced by a plasma is equal to the power input to maintain the plasma. This requires a plasma that is hot, dense, and well confined. The temperature required, about 100 million kelvins, is several times that of the Sun. The product of the density and energy confinement time of the plasma (the time it takes the plasma to lose its energy if unreplaced) must exceed a critical value.

Principal approaches to controlled thermonuclear fusion

There are two main approaches to controlled fusion—namely, magnetic confinement and inertial confinement. In magnetic confinement, a low-density plasma is confined for a long period of time by a magnetic field. The plasma density is roughly 10^{15} particles per cubic centimetre, which is many thousands of times less than the density of air at room temperature. The energy confinement time must then be at least one second—*i.e.*, the energy in the plasma must be replaced every second. In inertial confinement, no attempt is made to confine the plasma beyond the time it takes the plasma to disassemble. The energy confinement time is simply the time it takes the fusing plasma to expand. Confined only by its own inertia, the plasma survives for only about one-billionth of a second (one nanosecond). Hence, breakeven in this scheme requires a very large density of particles, typically about 10^{24} particles per cubic centimetre, which is about 100 times the density of a liquid. The extremely high density is achieved by compressing a solid pellet of fuel by the pressure of incident laser or particle beams. These approaches are sometimes referred to as laser fusion or particle-beam fusion.

The fusion reaction least difficult to achieve combines a deuteron (the nucleus of the deuterium atom) with a triton (the nucleus of a tritium atom). Both nuclei are isotopes of the hydrogen nucleus and contain a single unit of positive electric charge. Deuterium–tritium (D–T) fusion thus requires the nuclei to have lower kinetic energy than is needed for the fusion of more highly charged, heavier nuclei. The two products of the reaction are an alpha particle (nucleus of the helium atom) at an energy of 3.5 million electron volts (MeV) and a neutron at an energy of 14.1 MeV. (One MeV is the energy equivalent of 10 billion kelvins.) The neutron, lacking electric charge, is not affected by electric or magnetic fields within the plasma and can escape the plasma to deposit its energy in a material, such as lithium, which can surround the plasma. The heat generated in the lithium blanket is then converted to electrical energy by conventional means, such as turbines. The electrically charged alpha particle collides with the deuterons and tritons (by their electrical interaction) and can be magnetically confined within the plasma. It thereby transfers its energy to the reacting nuclei. When this redeposition of the fusion energy into the plasma exceeds the power lost from the plasma (by electromagnetic radiation, conduction, and convection), the plasma will be self-sustaining, or "ignited."

Inexhaustible source of energy

With deuterium and tritium as the fuel, the fusion reactor would be an effectively inexhaustible source of energy. Deuterium is obtained from seawater. About one in every 3,000 water molecules contains a deuterium atom. There is enough deuterium in the oceans to provide for the world's energy needs for billions of years. One gram of fusion fuel can produce as much energy as 9,000 litres of oil. The amount of deuterium found naturally in one litre of water is the energy equivalent of 300 litres of gasoline. Tritium is bred in the fusion reactor. It is generated in the lithium blanket as a product of the reaction in which neutrons are captured by the lithium nuclei.

A fusion reactor would have several attractive safety features. First, it is not subject to a runaway, or "meltdown," accident as is a fission reactor. The fusion reaction is not a chain reaction. It requires a hot plasma. Accidental interruption of a plasma control system would extinguish

the plasma and terminate fusion. Second, fusion reaction does not produce long-term radioactive wastes. Neutron bombardment will, however, activate the walls of the containment vessel. Such activation can be greatly reduced by employing fusion reactions that do not produce neutrons. Such "advanced" fusion-fuel cycles, as, for example, the fusion of deuterons and helium-3 nuclei, require higher temperatures than D–T fusion. Nearly neutron-free fusion systems might make up a "second generation" of fusion reactors. Finally, a fusion reactor would not release the gaseous pollutants that accompany the combustion of fossil fuels.

Principles of magnetic confinement. *Confinement physics.* Magnetic confinement of plasmas is the most highly developed approach to controlled fusion. The hot plasma is contained by magnetic forces exerted on the charged particles. A large part of the problem of fusion has been the attainment of magnetic field configurations that effectively confine the plasma. A successful configuration must meet three criteria: (1) the plasma must be in a time-independent equilibrium state, (2) the equilibrium must be macroscopically stable, and (3) the leakage of plasma energy to the bounding wall must be small.

Basic requirements

A single charged particle tends to spiral about a magnetic line of force. It is necessary that the single particle trajectories do not intersect the wall. Moreover, the pressure force, arising from the thermal energy of all the particles, is in a direction to expand the plasma. For the plasma to be in equilibrium, the magnetic force acting on the electric current within the plasma must balance the pressure force at every point in the plasma.

The equilibrium thus obtained has to be stable. A plasma is stable if after a small perturbation it returns to its original state. A plasma is continually perturbed by random thermal "noise" fluctuations. If unstable, the plasma might depart from its equilibrium state and rapidly escape the confines of the magnetic field (perhaps in less than one-thousandth of a second).

A plasma in stable equilibrium can be maintained indefinitely if the leakage of energy from the plasma is balanced by energy input, much like a leaky bucket can maintain a constant water level if the amount of water flowing in equals that flowing out. If the plasma energy loss is too large, then ignition cannot be achieved. An unavoidable diffusion of energy across the magnetic field lines will occur from the collisions between the particles. A collision can give the particle a random kick across the field. The net effect is to transport energy from the hot core to the wall. This transport process, known as classical diffusion, is theoretically not strong in hot fusion plasmas and is easily compensated for by heat from the alpha particle fusion products. In experiments, however, energy is lost from plasma more rapidly than expected from classical diffusion.

The observed energy loss typically exceeds the classical value by a factor of 10–100. Reduction of this anomalous transport is important to the engineering feasibility of fusion. An understanding of anomalous transport in plasmas in terms of physics is not yet in hand. A viewpoint under investigation is that the anomalous loss is caused by fine-scale turbulence in the plasma. Turbulently fluctuating electric and magnetic fields can push particles across the confining magnetic field. Solution of the anomalous transport problem involves research into fundamental topics in plasma physics, such as plasma turbulence.

Types of magnetic configurations

Many different types of magnetic configurations for plasma confinement have been devised and tested over the years. This development has been an evolutionary process that has resulted in a family of related magnetic configurations, which may be grouped into two classes: closed, toroidal configurations; and open, linear configurations. Toroidal devices are the most highly developed. A simple straight magnetic field would not suffice, since the plasma would be free to stream out the ends. The ends, and the end loss, can be eliminated by forming the plasma and magnetic field in the closed shape of a doughnut, or torus. In the approach involving linear devices, also called mirror confinement, the end loss is reduced by "plugging" the ends of such a device magnetically and electrostatically.

Toroidal confinement. The most extensively investigated toroidal confinement concept is the tokamak (Figure 87). The tokamak (an acronym derived from the Russian words for toroidal magnetic confinement) was introduced in the mid-1960s by Soviet plasma physicists. The magnetic lines of force are helixes that spiral around the torus. The helical magnetic field is composed of two components: (1) a toroidal magnetic field component, which points the long way around the torus; and (2) a poloidal field, which is directed the short way around the machine. The toroidal field is produced by current-carrying coils that surround the toroidal vacuum chamber containing the plasma. (The plasma must be situated within an evacuated chamber to prevent it from being cooled by interactions with air molecules.) The poloidal magnetic field is generated by a toroidal electric current that is forced to flow within the conducting plasma. Both magnetic field components are necessary for the plasma to be in stable equilibrium. If the poloidal field were zero so that the field lines were simply circles wrapped about the torus, then the plasma would not be in equilibrium. The particles would not strictly follow the field lines but would drift to the walls. The addition of the poloidal field provides particle orbits that are contained within the device. If the toroidal field were zero so that the magnetic field lines were directed only the short way around the torus, the plasma would be in equilibrium, but it would be unstable. The plasma column would develop growing distortions, or kinks, which would carry the plasma into the wall.

The tokamak

From R.W. Conn, "The Engineering of Magnetic Fusion Reactors," copyright © 1983 by Scientific American, Inc., all rights reserved

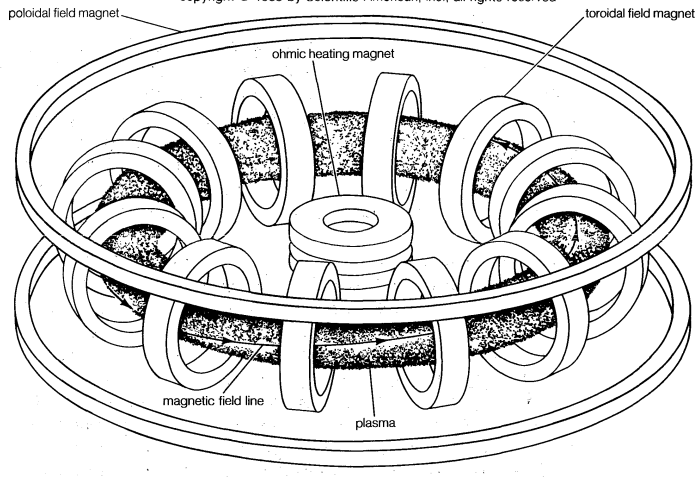


Figure 87: Tokamak magnetic confinement.

Several novel methods have been developed to drive the steady-state current that produces the poloidal magnetic field. A technique known as radio-frequency (RF) current drive employs electromagnetic waves to generate the current in a manner similar to the plasma-heating methods described below. Electromagnetic waves are injected into the plasma so that they propagate within the plasma in one direction around the torus. The speed of the waves is chosen to equal roughly the average speed of the electrons in the plasma. The wave electric field (which in a plasma has a component along its direction of travel) can then continuously accelerate the electrons as the wave and particles move together around the torus. In this way, the directed wave momentum is transferred to the electrons. The electrons develop a net motion, or current, in one direction. Although this technique is now well established, its efficiency is reduced at the density of a reacting plasma.

RF current drive

Another established current-drive technique is neutral-beam current drive. A beam of high-energy neutral atoms is injected into the plasma along the toroidal direction. The neutral beam will freely enter the plasma since it is unaffected by the magnetic field. Once inside the plasma, the neutral atoms of the beam become ionized by collisions with the electrons. The beam then consists of energetic positively charged nuclei (protons in the case of a hydrogen beam) that are confined within the plasma by the magnetic field. The high-speed ions travel toroidally

Neutral-beam current drive

along the magnetic field and collide with the electrons. The energy transfer from beam ion to electron will push the electrons in one direction and thereby produce a current. This technique is limited by the pace of the technological development of intense neutral-beam sources.

Faraday induction, or "ohmic current drive," can be used to initiate and build up the current. A magnetic flux that increases over time is produced through the hole in the torus. The plasma surrounds the flux. The time-varying flux induces a toroidal electric field that drives the plasma current. This technique efficiently drives a pulsed plasma current; however, it cannot be used for a steady-state current, which would require a magnetic flux increasing indefinitely over time. A pulsed reactor would suffer from engineering problems such as materials fatigue.

The plasma in a tokamak fusion reactor would have a major diameter in the range of 10 metres and a minor diameter in the range of roughly three metres. The plasma current would likely be tens of millions of amperes and the toroidal magnetic field would be several teslas. The coils that produce the strong toroidal magnetic field would probably be superconducting in order to minimize the power dissipation in the coils.

Other toroidal confinement concepts that offer potential advantages over the tokamak are under development, albeit at a smaller scale than tokamak research. Two such alternatives under investigation at various laboratories across the world are the reversed-field pinch (RFP) and stellarator concepts. Both configurations use helical magnetic fields similar to those of the tokamak. The reversed-field pinch operates with a low toroidal magnetic field. This results in a compact, high-power density reactor with ordinary (instead of superconducting) coils. The stellarator produces its magnetic field by external helical coils only. The plasma current is nearly zero, and plasma current drive is not required.

Reversed-field pinch and stellarator

Mirror confinement. An alternative approach to magnetic confinement is to employ a straight configuration in which the end loss is reduced by a combination of magnetic and electric plugging. In such a linear fusion reactor the magnetic field strength is increased at the ends. Charged particles that approach the end slow down, and many are reflected from this "magnetic mirror." The same magnetic reflection mechanism traps particles in the Earth's magnetosphere (specifically in the Van Allen belts). Unfortunately, particles with extremely high speed along the field are not stopped by the mirror. To inhibit this leakage out of the ends, electrostatic plugging is provided. An additional section of plasma is added at each end beyond the magnetic mirror. The plasma in these "end plugs" is of sufficient density and temperature to produce an electrostatic potential barrier to nuclei. The electric force on the nuclei would stop much of the leakage through the magnetic mirror. The final configuration is called a tandem mirror.

Plasma heating. A fusion reactor requires tens of megawatts of heating to reach ignition temperature. Two plasma-heating methods have been highly developed: electromagnetic wave heating and neutral-beam injection heating. In the former, electromagnetic waves are launched by antennas at the surface of the plasma. The waves penetrate the plasma and transfer their energy to the constituent particles. Ionized gases can support the propagation of a remarkably large variety of waves not found in other forms of matter. Effective wave-heating techniques employ frequencies from the radio-frequency range (tens of megahertz) to the microwave range (tens of gigahertz). Power sources are available over such frequencies. Power absorption often relies upon a resonant interaction between the wave and plasma. For example, if the frequency of the electromagnetic wave is equal to the frequency at which a nucleus gyrates about a magnetic field line, this resonant nucleus absorbs energy from the wave. This technique is called ion cyclotron resonance heating. Similarly, electron cyclotron resonance heating may be used to heat electrons. Such electron heating requires very high frequency (tens to hundreds of gigahertz). Recently developed free-electron lasers and gyrotron tubes are required at the highest frequencies.

Electromagnetic wave heating

In the second method, beams of neutral atoms at high energy (up to about one million electron volts) are injected into the plasma. The approach is quite similar to the neutral-beam current drive that was described above. When used for heating, however, the beams are injected in both directions around the torus, so that no net momentum is imparted to the plasma. The slowing down, or transfer of beam energy to the plasma, constitutes the heating mechanism.

Neutral beam injection heating

Principles of inertial confinement. In an inertial confinement fusion (ICF) reactor a tiny solid pellet of fuel (such as deuterium–tritium) would be compressed to tremendous density and temperature so that fusion power is produced in the very short time (tens of nanoseconds) before the pellet blows apart. The compression is accomplished by focusing an intense laser beam (or a charged particle beam) upon the small pellet (typically one to 10 millimetres in diameter). The surface of the pellet is ionized by the beam. The ablation of the ionized material generates a large inward force on the pellet (as in the rocket effect), which compresses the pellet to 1,000 to 10,000 times liquid density. During compression the temperature of the pellet increases to a value sufficient to produce fusion reactions (Figure 88). Ignition occurs, and the pellet, now a dense plasma, is burned up in a small micro-explosion. The process is repeated between one and 100 times per second.

Laser fusion

From F.F. Chen, *Introduction to Plasma Physics* (1974), Plenum Press

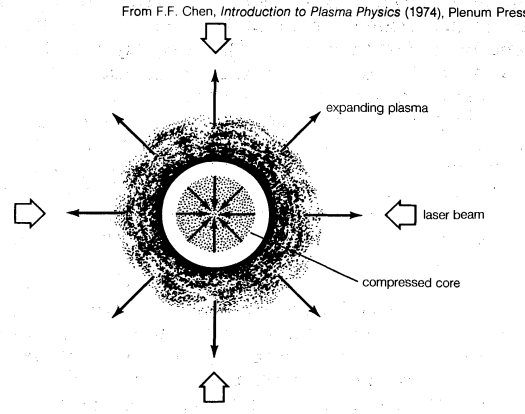


Figure 88: Laser fusion.

Inertial confinement fusion has been compared to the four-stage internal-combustion engine. In the fuel-injection stage, the pellet is injected into a blast chamber. In the compression stage, the pellet is compressed by the driver beams (laser or charged particle). In the ignition stage, the fusion reactions begin. In the final stage, the burn proceeds to completion and the fusion reaction products (neutrons, X rays, and charged particles) bombard and heat a blanket.

For efficient thermonuclear burn, the time to burn the pellet must be less than the disassembly time. This yields a criterion that in the compressed state the product of the pellet mass density and the pellet radius exceed about three grams per square centimetre. A high mass density will hasten the burn, and a large radius will slow the disassembly time. This criterion can be satisfied, for example, with a one-millimetre pellet and a fuel density of 30 grams per cubic centimetre. This density requires pellet compression to about 150 times its initial density (4.5×10^{22} particles per cubic centimetre).

The two key components of an ICF reactor are the driver beams and pellets. The power requirement of the driver beams is influenced by the efficiency of the driver (conversion of electrical power to beam power) and the efficiency at which the driver energy is absorbed by the pellet. At present, each of these efficiencies is between 1 and 10 percent for laser fusion. To overcome these losses requires an energy gain Q (fusion power/absorbed power) of about 1,000 to 10,000. An energy of one to five megajoules must be delivered to the pellet in one to 10 nanoseconds. Thus, the laser power is enormous, about one trillion watts. The laser irradiation must strike the pellet surface uniformly to compress the pellet effectively. Thus, many laser beams

irradiate the pellet with approximate spherical symmetry (see Figure 88).

Particle beam fusion

Charged particle driver beams can offer the advantage of more efficient production and absorption. Beams of electrons would "defocus" as a result of the small electron mass. Thus, light ion or heavy ion beams are employed. Beams of light ions (hydrogen through carbon) in the MeV energy range are produced by pulsed-power diode accelerators. In such accelerators, large voltage between a cathode and anode applies a strong electric force on the particles. The expected efficiency of light ion production is about 30 percent. Beams of heavy ions (xenon to uranium) are accelerated to energies of one billion electron volts (GeV) using accelerator technology from high-energy physics experiments. Such accelerators push particles by induction or with electromagnetic waves.

The pellets that are currently used are multilayered, consisting of several concentric spheres. This design has been chosen to optimize the symmetry, stability, and isentropy of the compression process (see below). The outermost layer is the ablation layer that absorbs the driver energy. It is converted to a plasma and blown away. In the plasma state, the high heat conduction symmetrizes the ablation process. The ablation layer surrounds a thick, heavy material of high atomic mass. The recoil from the ablation implodes the heavy layer, producing a shock wave that compresses and heats the next inner layer of deuterium–tritium fuel. The implosion speed is several hundred kilometres per second, produced by a force equivalent to some 10 billion atmospheres. The heavy layer inhibits the growth of instabilities (nonsymmetrical distortions) that would interfere with the compression. It also prevents the high-energy electrons and protons from the corona plasma from heating the deuterium–tritium layer. In this way, the compression is isentropic. The laser energy provides compression, not heat (entropy). The burn initiates in the D–T layer and spreads outward as the alpha particles collide with and heat the rest of the pellet to thermonuclear temperature.

Development of fusion reactor technology. *Magnetic confinement.* Several decades of fusion research have produced accomplishments of two types. First, the discipline of plasma physics has developed to the point that theoretical and experimental tools permit quantitative evaluation of many aspects of fusion reactor concepts. Second, and perhaps most revealing, the evolutionary improvement of plasma parameters has placed experiments at the threshold of energy breakeven.

Fusion research experiments are usually performed with hydrogen or deuterium plasmas. Radioactive tritium is not added since remote-handling requirements would unnecessarily complicate the experiments. A figure of merit with which to judge the plasma quality is the energy gain Q (= fusion power/heating power) that would occur if the plasma contained tritium. Over the past few decades Q has increased by more than one million, from 10^{-7} (less than one-millionth) in 1965 to about 0.5 in 1989. It is now a virtual certainty that the long-sought goal of energy breakeven $Q = 1$, is achievable. Current experiments confine plasmas with volumes of 100 cubic metres at temperatures in excess of 100 million kelvins (up to 20 kiloelectron volts) and with energy confinement times of a fraction of a second.

There is a wide variety of plasma experiments under way to investigate many aspects of the fusion problem. Parameters closest to the reactor level have been attained in the four large "flagship" experiments in the United States, western Europe, Japan, and the Soviet Union. These large tokamak facilities are the Tokamak Fusion Test Reactor (TFTR) at the Princeton Plasma Physics Laboratory in the United States; the Joint European Torus (JET), a multinational western European venture operated in Britain; JT-60 of the Japan Atomic Energy Research Institute; and T-15 at the I.V. Kurchatov Institute of Atomic Energy in Moscow.

A next major step in the development of fusion power is the construction of a facility to study the physics of a burning, ignited plasma (with Q being infinite). The presence of alpha particles can alter the behaviour of the plasma in ways not easily simulated in nonburning plasmas. A design effort is presently under consideration in the United States for an ignited-plasma facility called the Compact Ignition Tokamak (CIT). The experiment would employ extremely high magnetic fields (about 10 teslas) to enable the device to be relatively compact and thereby cost-effective. In recent years a consensus has emerged that the technical know-how exists to construct such an ignition device.

The next step after an ignition physics experiment would be a test reactor that would provide an integrated reactor facility, including materials and plasma engineering aspects. Under the auspices of the International Atomic Energy Agency, an international team is presently exploring a conceptual design of such a device, known as the International Thermonuclear Experimental Reactor (ITER). The expense of the device encourages collaboration between the United States, the countries of western Europe, the Soviet Union, and Japan. It is conceived that a device such as ITER would be followed by a demonstration fusion reactor power plant.

With the tremendous advances in scientific understanding and plasma quality, questions regarding the engineering and economic attractiveness of the tokamak concept have received greater attention. Materials development is required. For example, the wall exposed to the plasma must survive intense neutron bombardment. The optimal path to fusion-energy production involves some balance between further upscaling of the current tokamak concept toward reactor parameters and improvement of the magnetic confinement concept. Improvements can accrue from enhanced scientific understanding through research and by the development of alternative concepts, such as the stellarator and the reversed-field pinch.

Inertial confinement. ICF research has followed an evolutionary path similar to that of magnetic fusion. In the laser fusion approach, densities of 100 to 200 times liquid deuterium–tritium density have been achieved. For example, at the Lawrence Livermore National Laboratory in California, a product of density and energy-confinement time of 5×10^{14} seconds per cubic centimetre has been achieved using the world's largest and most powerful laser, the so-called Nova laser. (The Nova is a 10-beam neodymium–glass laser operated at an energy of 20,000 joules in a one-nanosecond pulse.) Although the value of this product is comparable to that representing breakeven for magnetic fusion, laser fusion requires a larger value to overcome the poor efficiency of existing lasers.

As a result of such progress, plans are being considered in the United States for a laboratory microfusion facility capable of demonstrating a fusion energy gain (fusion power/absorbed driver power) of 100 with a fusion energy release of nearly one billion joules. Light-ion beam experiments are also making headway. Scientists at the Particle Beam Fusion Accelerator (PBFA) at Sandia National Laboratories in New Mexico have demonstrated the ability to focus 72 beams to a spot diameter of less than six millimetres. One objective is to achieve an intensity of 10 terawatts per cubic centimetre (100 terawatts per cubic centimetre are required in a practical fusion reactor).

In both the magnetic and inertial confinement programs, the experimental steps become increasingly more expensive as the reactor regime is approached. At the same time, basic research and innovation are needed to enhance the attractiveness of the reactor concepts. Significant wisdom is required to balance these needs and to build effectively upon the impressive results to date so that nuclear fusion can indeed become a major factor in meeting the world's ever-growing energy needs. (S.C.P.)

BIBLIOGRAPHY

The concept of energy: General introductions are provided by RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* 3 vol. (1963–65; vol. 1 and 2 have been reprinted, 1977); MITCHELL WILSON, *Energy,* rev. ed. (1970); *Energy: Readings from Scientific American,* with introductions by S. FRED SINGER (1979); and JANET RAMAGE, *Energy: A Guidebook* (1983).

History of energy-conversion technology: Historical developments are outlined in CHARLES SINGER *et al.* (eds.), *A History*

of Technology, 8 vol. (1954–84); MAURICE DAUMAS (ed.), *Histoire général des techniques,* 5 vol. (1962–79)—the first 3 vol. have been translated as *A History of Technology and Invention: Progress Through the Ages* (1969–79); and MELVIN KRANZBERG and CARROLL W. PURSELL, JR. (eds.), *Technology in Western Civilization,* 2 vol. (1967). (Ed.)

Major energy-conversion devices and systems: (*Turbines*): General principles are considered in AUBREY F. BURSTALL, *A History of Mechanical Engineering* (1963); G.T. CSANADY, *Theory of Turbomachines* (1964); and CALVIN VICTOR DAVIS and KENNETH E. SORENSEN (eds.), *Handbook of Applied Hydraulics,* 3rd ed. (1969, reprinted 1984). Discussions of steam and wind turbines are provided by W.G. STELTZ and A.M. DONALDSON (eds.), *Aero-thermodynamics of Steam Turbines* (1981); and GARY L. JOHNSON, *Wind Energy Systems* (1985). (Fr.L.)

(*Internal-combustion engines*): A historical treatment of the invention of the internal-combustion engine is provided in two articles in *Technology and Culture* by LYNWOOD BRYANT, "The Silent Otto," 7(2):184–200 (Spring 1966), and "The Origin of the Four-Stroke Cycle," 8(2):178–198 (April 1967). Overviews can be found in LESTER C. LICHTY, *Combustion Engine Processes* (1967); EDWARD F. OBERT, *Internal Combustion Engines and Air Pollution* (1973); ASHLEY S. CAMPBELL, *Thermodynamic Analysis of Combustion Engines* (1979, reprinted 1985); CHARLES FAYETTE TAYLOR, *The Internal-Combustion Engine in Theory and Practice,* 2nd ed. rev., 2 vol. (1985); COLIN R. FERGUSON, *Internal Combustion Engines: Applied Thermosciences* (1986); and JOHN B. HEYWOOD, *Internal Combustion Engine Fundamentals* (1988).

Diesel engines are discussed in S.D. HADDAD and N. WATSON (eds.), *Principles and Performance in Diesel Engineering* (1984); and FRANK J. THIESSEN and DAVIS N. DALES, *Diesel Fundamentals,* 2nd ed. (1986). (C.L.P.II)

For gas-turbine engines, see WILLIAM W. BATHIE, *Fundamentals of Gas Turbines* (1984); and FRANK WHITTLE, *Gas Turbine Aero-thermodynamics: With Special Reference to Aircraft Propulsion* (1981). (Ed.)

Texts on jet engines include two nontechnical works, ROLLS-ROYCE LTD., *The Jet Engine,* 4th ed. (1986), with a discussion of basic concepts and a systematic analysis of jet engine components; and IRWIN E. TREAGER, *Aircraft Gas Turbine Engine Technology,* 2nd ed. (1979), with a section on the history of the jet engine. For more technical treatments, see JACK L. KERREBROCK, *Aircraft Engines and Gas Turbines* (1977), which deals primarily with the thermodynamic and aerodynamic operation of major engine components; and GORDON C. OATES, *Aerothermodynamics of Gas Turbine and Rocket Propulsion,* rev. and enlarged ed. (1988). (F.F.E.)

WERNHER VON BRAUN and FREDERICK I. ORDWAY III, *Space Travel: A History,* 4th ed. rev. in collaboration with DAVID DOOLING (1985); and WILLY LEY, *Rockets, Missiles, and Space Travel,* rev. and enlarged ed. (1961), offer introductions to the history of rocketry. Rocket engines are discussed in MARCEL BARRÈRE *et al., Rocket Propulsion* (1960; originally published in French, 1957); and GEORGE P. SUTTON, *Rocket Propulsion Elements: An Introduction to the Engineering of Rockets,* 5th ed. (1986). (E.W.P.)

(*Nuclear fission reactors*): RICHARD RHODES, *The Making of the Atomic Bomb* (1986), is a history of developments leading to the first reactor and first atomic bomb. An elementary text covering reactor concepts, radiation, nuclear fuel cycles, reactor systems, safety and safeguards, and fusion concepts is RONALD ALLEN KNIEF, *Nuclear Energy Technology: Theory and Practice of Commercial Nuclear Power* (1981); the same concepts are treated at a more advanced mathematical level in JOHN R. LAMARSH, *Introduction to Nuclear Engineering,* 2nd ed. (1983). JAMES J. DUDERSTADT and LOUIS J. HAMILTON, *Nuclear Reactor Analysis* (1976), discusses the theory of neutron behaviour in matter, criticality, neutron spectrum, and reactor core design and control, with emphasis on methods of calculation. MANSON BENEDICT, THOMAS H. PIGFORD, and HANS WOLFGANG LEVI, *Nuclear Chemical Engineering,* 2nd ed. (1981), includes coverage of fuel cycles, the chemistry of uranium and heavy elements, the theory of multistage systems, enrichment processes and theory, the reprocessing of nuclear fuel, and nuclear waste management. Current developments in domestic and international nuclear power, safety, research, and opinion are published in *Nuclear News* (monthly), the newsletter of the American Nuclear Society. (B.I.S.)

(*Electric generators and electric motors*): Overviews may be found in SYED A. NASAR (ed.), *Handbook of Electric Machines* (1987); G.R. SLEMON and A. STRAUGHEN, *Electric Machines* (1980); SYED A. NASAR and L.E. UNNEWEHR, *Electromechanics and Electric Machines,* 2nd ed. (1983); VINCENT DEL TORO, *Electric Machines and Power Systems* (1985); and GEORGE MCPHERSON and ROBERT D. LARAMORE, *An Introduction to Electrical Machines and Transformers,* 2nd ed. (1990). (G.R.Sl.)

Direct energy-conversion devices: STANLEY W. ANGRIST, *Direct Energy Conversion,* 4th ed. (1987), provides a historical introduction and overview of the devices discussed below.

(*Batteries and fuel cells*): Overviews include COLIN A. VINCENT *et al., Modern Batteries: An Introduction to Electrochemical Power Sources* (1984), written for the nonspecialist; MANFRED BREITER, *Electrochemical Processes in Fuel Cells* (1969); and ROBERT NOYES (ed.), *Fuel Cells for Public Utility and Industrial Power* (1977). DAVID LINDEN (ed.), *Handbook of Batteries and Fuel Cells* (1984), provides comprehensive information on types and applications. (B.S.)

(*Solar cells*): PAUL D. MAYCOCK and EDWARD N. STIREWALT, *Photovoltaics: Sunlight to Electricity in One Step* (1981), is a nontechnical work. RICHARD J. KOMP, *Practical Photovoltaics: Electricity from Solar Cells,* 2nd ed. (1984); and KENNETH ZWEIBEL and PAUL HERSCH, *Basic Photovoltaic Principles and Methods* (1984), are more advanced but still accessible to the nontechnically trained reader. STEPHEN J. FONASH, *Solar Cell Device Physics* (1981), is for the specialist. (S.J.F./R.T.F)

(*Thermoelectric power generators*): General references include A.F. IOFFE, *Semiconductor Thermoelements, and Thermoelectric Cooling* (1957; originally published in Russian, 1956), two classic works emphasizing the important contributions made at the Institute for Semiconductors in Leningrad; H.J. GOLDSMID, *Applications of Thermoelectricity* (1960), a brief readable monograph covering thermoelectric effects, materials, devices, and applications; and ROBERT R. HEIKES and ROLAND W. URE, JR., *Thermoelectricity: Science and Engineering* (1961), a review of all aspects of thermoelectric devices. See also J.W.C. HARPSTER, P.R. SWINEHART, and F. BRAUN, "Solid State Thermal Control for Spacecraft," *Solid-State Electronics,* 18(6):551–555 (June 1975), an examination of the heat-pumping capabilities of thermoelectric devices in Earth-orbiting spacecraft. (J.W.H.)

(*Thermionic power converters*): Texts on thermodynamics in general include LEIGHTON E. SISSOM and DONALD R. PITTS, *Elements of Transport Phenomena* (1972); and FRANCIS F. HUANG, *Engineering Thermodynamics: Fundamentals and Applications,* 2nd ed. (1988). Discussions on thermionic converters in particular are G.N. HATSOPOULOS and E.P. GYFTOPOULOS, *Thermionic Energy Conversion,* 2 vol. (1973–79); and F.G. BAKSHT *et al., Thermionic Converters and Low-Temperature Plasma,* trans. from Russian (1978). (L.E.Si.)

(*Magnetohydrodynamic power generators*): RICHARD J. ROSA, *Magnetohydrodynamic Energy Conversion* (1968, reprinted 1987); GEORGE W. SUTTON and ARTHUR SHERMAN, *Engineering Magnetohydrodynamics* (1965); and V.A. KIRILLIN and A.E. SCHEINDLIN (eds.), *MHD Energy Conversion: Physiotechnical Problems* (1986; originally published in Russian, 1983), are general texts on MHD principles and applications. Useful journal articles include three from *Magnetohydrodynamics: An International Journal,* vol. 2, no. 1 (1989): L.H.TH. RIETJENS, "MHD for Large-Scale Electrical Power Generation in the 21st Century," pp. 17–25; E.P. VELIKHOV *et al.,* "Pulsed MHD Facilities: Geophysical Applications," pp. 27–33; and A.E. SCHEINDLIN and W.D. JACKSON, "Ninth International Conference on Magnetohydrodynamic Electrical Power Generation: Status Report Summary," pp. 11–16. Open-cycle MHD is treated in J.B. HEYWOOD and G.J. WOMACK (eds.), *Open-Cycle MHD Power Generation* (1969); and M. PETRICK and B. YA. SHUMYATSKY, *Open-Cycle Magnetohydrodynamic Electrical Power Generation* (1978), a joint U.S.–U.S.S.R. publication. Two conference proceedings are among important sources of current information: papers from meetings of the SYMPOSIUM ON THE ENGINEERING ASPECTS OF MAGNETOHYDRODYNAMICS, an American conference; and from the series of meetings of the INTERNATIONAL CONFERENCE ON MHD ELECTRICAL POWER GENERATION. (W.D.J.)

(*Fusion reactors*): Articles written for the lay reader include two from *Scientific American:* ROBERT W. CONN, "The Engineering of Magnetic Fusion Reactors," 249(4):60–71 (October 1983); and R. STEPHEN CRAXTON, ROBERT L. MCCRORY, and JOHN M. SOURES, "Progress in Laser Fusion," 255(2):68–79 (August 1986). The following books assume that the reader has a science background. Concepts of fusion in general are examined by THOMAS JAMES DOLAN, *Fusion Research: Principles, Experiments, and Technology* (1982). FRANCIS F. CHEN, *Introduction to Plasma Physics and Controlled Fusion,* vol. 1, *Plasma Physics,* 2nd ed. (1984); and WESTON M. STACEY, JR., *Fusion Plasma Analysis* (1981), provide introductions to plasma physics. Particular approaches to fusion are analyzed in JAMES J. DUDERSTADT and GREGORY A. MOSES, *Inertial Confinement Fusion* (1982); and WESTON M. STACEY, JR., *Fusion: An Introduction to the Physics and Technology of Magnetic Confinement Fusion* (1984). (S.C.P.)

Engineering

Engineering is the professional art of applying science to the optimum conversion of the resources of nature to the uses of humankind. Engineering has been defined by the Engineers Council for Professional Development, in the United States, as the creative application of "scientific principles to design or develop structures, machines, apparatus, or manufacturing processes, or works utilizing them singly or in combination; or to construct or operate the same with full cognizance of their design; or to forecast their behaviour under specific operating conditions; all as respects an intended function, economics of operation and safety to life and property." The term engineering is sometimes more loosely defined, especially in Great Britain, as the manufacture or assembly of engines, machine tools, and machine parts.

The words engine and ingenious are derived from the same Latin root, *ingenerare,* which means "to create." The early English verb *engine* meant "to contrive." Thus the engines of war were devices such as catapults, floating bridges, and assault towers; their designer was the "engine-er," or military engineer. The counterpart of the military engineer was the civil engineer, who applied essentially the same knowledge and skills to designing buildings, streets, water supplies, sewage systems, and other projects.

Associated with engineering is a great body of special knowledge; preparation for professional practice involves extensive training in the application of that knowledge. Standards of engineering practice are maintained through the efforts of professional societies, usually organized on a national or regional basis, with each member acknowledging a responsibility to the public over and above responsibilities to his employer or to other members of his society.

The function of the scientist is to know, while that of the engineer is to do. The scientist adds to the store of verified, systematized knowledge of the physical world; the engineer brings this knowledge to bear on practical problems. Engineering is based principally on physics, chemistry, and mathematics and their extensions into materials science, solid and fluid mechanics, thermodynamics, transfer and rate processes, and systems analysis.

Unlike the scientist, the engineer is not free to select the problem that interests him; he must solve problems as they arise; his solution must satisfy conflicting requirements. Usually efficiency costs money; safety adds to complexity; improved performance increases weight. The engineering solution is the optimum solution, the end result that, taking many factors into account, is most desirable. It may be the most reliable within a given weight limit, the simplest that will satisfy certain safety requirements, or the most efficient for a given cost. In many engineering problems the social costs are significant.

Engineers employ two types of natural resources—materials and energy. Materials are useful because of their properties: their strength, ease of fabrication, lightness, or durability; their ability to insulate or conduct; their chemical, electrical, or acoustical properties. Important sources of energy include fossil fuels (coal, petroleum, gas), wind, sunlight, falling water, and nuclear fission. Since most resources are limited, the engineer must concern himself with the continual development of new resources as well as the efficient utilization of existing ones.

For the history and functions of industrial engineering, see INDUSTRIAL ENGINEERING AND PRODUCT MANAGEMENT.

The article is divided into the following sections:

Engineering as a profession

HISTORY OF ENGINEERING

The first engineer known by name and achievement is Imhotep, builder of the Step Pyramid at Ṣaqqārah, Egypt, probably in about 2550 BC. Imhotep's successors—Egyptian, Persian, Greek, and Roman—carried civil engineering to remarkable heights on the basis of empirical methods aided by arithmetic, geometry, and a smattering of physical science. The Pharos (lighthouse) of Alexandria, Solomon's Temple in Jerusalem, the Colosseum in Rome, the Persian and Roman road systems, the Pont du Gard aqueduct in France, and many other large structures, some of which endure to this day, testify to their skill, imagination, and daring. Of many treatises written by them, one in particular survives to provide a picture of engineering education and practice in classical times: Vitruvius' *De architectura,* published in Rome in the 1st century AD, a 10-volume work covering building materials, construction methods, hydraulics, measurement, and town planning.

In construction medieval European engineers carried technique, in the form of the Gothic arch and flying buttress, to a height unknown to the Romans. The sketchbook of the 13th-century French engineer Villard de Honnecourt reveals a wide knowledge of mathematics, geometry, natural and physical science, and draftsmanship.

In Asia, engineering had a separate but very similar development, with more and more sophisticated techniques of construction, hydraulics, and metallurgy helping to create advanced civilizations such as the Mongol empire, whose large, beautiful cities impressed Marco Polo in the 13th century.

Civil engineering emerged as a separate discipline in the 18th century, when the first professional societies and

Civil and mechanical engineering

schools of engineering were founded. Civil engineers of the 19th century built structures of all kinds, designed water-supply and sanitation systems, laid out railroad and highway networks, and planned cities. England and Scotland were the birthplace of mechanical engineering, as a derivation of the inventions of the Scottish engineer James Watt and the textile machinists of the Industrial Revolution. The development of the British machine-tool industry gave tremendous impetus to the study of mechanical engineering both in Britain and abroad.

The growth of knowledge of electricity—from Alessandro Volta's original electric cell of 1800 through the experiments of Michael Faraday and others, culminating in 1872 in the Gramme dynamo and electric motor (named after the Belgian Z.T. Gramme)—led to the development of electrical and electronics engineering. The electronics aspect became prominent through the work of such scientists as James Clerk Maxwell of Britain and Heinrich Hertz of Germany in the late 19th century. Major advances came with the development of the vacuum tube by Lee De Forest of the United States in the early 20th century and the invention of the transistor in the mid-20th century. In the late 20th century electrical and electronics engineers outnumbered all others in the world.

Chemical engineering

Chemical engineering grew out of the 19th-century proliferation of industrial processes involving chemical reactions in metallurgy, food, textiles, and many other areas. By 1880 the use of chemicals in manufacturing had created an industry whose function was the mass production of chemicals. The design and operation of the plants of this industry became a function of the chemical engineer.

ENGINEERING FUNCTIONS

Problem solving is common to all engineering work. The problem may involve quantitative or qualitative factors; it may be physical or economic; it may require abstract mathematics or common sense. Of great importance is the process of creative synthesis or design, putting ideas together to create a new and optimum solution.

Although engineering problems vary in scope and complexity, the same general approach is applicable. First comes an analysis of the situation and a preliminary decision on a plan of attack. In line with this plan, the problem is reduced to a more categorical question that can be clearly stated. The stated question is then answered by deductive reasoning from known principles or by creative synthesis, as in a new design. The answer or design is always checked for accuracy and adequacy. Finally, the results for the simplified problem are interpreted in terms of the original problem and reported in an appropriate form.

In order of decreasing emphasis on science, the major functions of all engineering branches are the following:

Research. Using mathematical and scientific concepts, experimental techniques, and inductive reasoning, the research engineer seeks new principles and processes.

Development. Development engineers apply the results of research to useful purposes. Creative application of new knowledge may result in a working model of a new electrical circuit, a chemical process, or an industrial machine.

Design. In designing a structure or a product, the engineer selects methods, specifies materials, and determines shapes to satisfy technical requirements and to meet performance specifications.

Construction. The construction engineer is responsible for preparing the site, determining procedures that will economically and safely yield the desired quality, directing the placement of materials, and organizing the personnel and equipment.

Production. Plant layout and equipment selection are the responsibility of the production engineer, who chooses processes and tools, integrates the flow of materials and components, and provides for testing and inspection.

Operation. The operating engineer controls machines, plants, and organizations providing power, transportation, and communication; determines procedures; and supervises personnel to obtain reliable and economic operation of complex equipment.

Management and other functions. In some countries and industries, engineers analyze customers' requirements, recommend units to satisfy needs economically, and resolve related problems. (R.J.Sm./Ed.)

Major fields of engineering

MILITARY ENGINEERING

In its earliest uses the term engineering referred particularly to the construction of engines of war and the execution of works intended to serve military purposes. Military engineers were long the only ones to whom the title engineer was applied.

The role of the military engineer in modern war is to apply engineering knowledge and resources to the furtherance of the commander's plans. The basic requirement is a sound general engineering knowledge directed to the technical aspects of those tasks likely to be encountered in war. Engineering work is influenced by topographical considerations and in battle also by tactical limitations. At times engineering factors will actually govern the choice of the military plan adopted; a military engineer must, therefore, possess a sound military education so that the best technical advice will be given to the commander.

History. In the prehistoric period every man was a fighter and every fighter was to some extent an engineer. Primitive efforts were restricted to the provision of artificial protection for the person and machines for hurling destruction at the enemy. In the earliest war annals it is difficult to distinguish the military from the civil engineer. Julius Caesar referred to his *praefectus fabrum,* an official who controlled the labour gangs employed on road making and also parties of artisans. The Domesday survey of AD 1086 included one "Waldivus Ingeniator," who held nine manors direct from the crown and was probably William the Conqueror's chief engineer in England. Throughout the Middle Ages, ecclesiastics were frequently employed as military engineers, not only for purposes of planning and building but also for fighting. One of the best known is Gundulph, bishop of Rochester, who built the White Tower of the Tower of London and Rochester Castle.

Thus, in ancient and medieval times the military engineer became a specialist who made and used engines of war such as catapults, ballistas, battering rams, ramps, towers, scaling ladders, and other devices in attacking or defending castles, fortresses, and fortified camps. In peacetime the military engineer built fortifications for the defense of the country or city. Because such engineers frequently dug trenches or tunnels as means of approaching or undermining enemy positions, they came to be called sappers or miners. With the invention of gunpowder and the countless other inventions that came in later centuries, the military engineer was required to have far more technical knowledge. He nevertheless remained a soldier and fought side by side with the infantry in many wars.

The French corps of engineers

Before the late 17th century the engineers of French armies were selected infantry officers given brevets as engineers; they performed both civil and military duties for the king's service. In 1673 Sébastien Le Prestre de Vauban was appointed director general of the royal fortifications, and it was largely owing to this great designer of fortified places that in 1690 an officer corps of engineers was established. Sapper and miner companies were formed later, although these units were generally attached to the artillery. In 1801 the officer corps of engineers was integrated with the sapper and miner units, and the amalgamated corps served with great distinction throughout Napoleon's campaigns. In 1868 military telegraphists were added to the corps. The first engineer railway battalion was formed in 1876, and a battalion of aeronauts raised in 1904 was the forerunner of the French air force.

The first military engineering school was established at Mézières in 1748, and Lazare Carnot, a former graduate of Mézières, moved the school to Metz in 1795, where it was renamed the École Polytechnique ("Polytechnic School").

Military engineering functions. The functions of modern military engineers vary among the armies of the world, but as a rule they include the following activities: (1) construction and maintenance of roads, bridges, airfields, landing strips, and zones for the airdrop of personnel and supplies, (2) interference with the enemy's mobility

by means of demolitions, floods, destruction of matériel, mine fields, and obstacles and fortifications of many types, (3) mapping and aiding the artillery to survey gun positions, rocket-launching sites, and target areas, (4) supplying water and engineering equipment, and (5) disposal of unexploded bombs or warheads. In the British army the Royal Engineers also operate the army postal service.

The U.S. Army Corps of Engineers is both a combatant arm and a technical service. Alone among the arms and services, it engages in civil as well as military activities. During the 20th century its civil works activities have centred upon the planning, construction, and maintenance of improvements to rivers, harbours, and other waterways and upon flood control. The principal military service performed by the Corps of Engineers in the United States and abroad is the construction and maintenance of buildings and utilities. In theatres of operation in wartime, such construction is carried out by engineer troops. In the United States in peace and war and overseas in peacetime, such construction is usually accomplished by private industry under contract to the Corps of Engineers. (Ed.)

CIVIL ENGINEERING

The term civil engineering was first used in the 18th century to distinguish the newly recognized profession from military engineering, until then preeminent. From earliest times, however, engineers have engaged in peaceful activities, and many of the civil engineering works of ancient and medieval times—such as the Roman public baths, roads, bridges, and aqueducts; the Flemish canals; the Dutch sea defenses; the French Gothic cathedrals; and many other monuments—reveal a history of inventive genius and persistent experimentation.

History. The beginnings of civil engineering as a separate discipline may be seen in the foundation in France in 1716 of the Bridge and Highway Corps, out of which in 1747 grew the École Nationale des Ponts et Chaussées ("National School of Bridges and Highways"). Its teachers wrote books that became standard works on the mechanics of materials, machines, and hydraulics, and leading British engineers learned French to read them. As design and calculation replaced rule of thumb and empirical formulas, and as expert knowledge was codified and formulated, the nonmilitary engineer moved to the front of the stage. Talented, if often self-taught, craftsmen, stonemasons, millwrights, toolmakers, and instrument makers became civil engineers. In Britain, James Brindley began as a millwright and became the foremost canal builder of the century; John Rennie was a millwright's apprentice who eventually built the new London Bridge; Thomas Telford, a stonemason, became Britain's leading road builder.

Smeaton's work

John Smeaton, the first man to call himself a civil engineer, began as an instrument maker. His design of Eddystone Lighthouse (1756–59), with its interlocking masonry, was based on a craftsman's experience. Smeaton's work was backed by thorough research, and his services were much in demand. In 1771 he founded the Society of Civil Engineers (now known as the Smeatonian Society). Its object was to bring together experienced engineers, entrepreneurs, and lawyers to promote the building of large public works, such as canals (and later railways), and to secure the parliamentary powers necessary to execute their schemes. Their meetings were held during parliamentary sessions; the society follows this custom to this day.

The École Polytechnique was founded in Paris in 1794, and the Bauakademie was started in Berlin in 1799, but no such schools existed in Great Britain for another two decades. It was this lack of opportunity for scientific study and for the exchange of experiences that led a group of young men in 1818 to found the Institution of Civil Engineers. The founders were keen to learn from one another and from their elders, and in 1820 they invited Thomas Telford, by then the dean of British civil engineers, to be their first president. There were similar developments elsewhere. By the mid-19th century there were civil engineering societies in many European countries and the United States, and the following century produced similar institutions in almost every country in the world.

Engineering education

Formal education in engineering science became widely available as other countries followed the lead of France and Germany. In Great Britain the universities, traditionally seats of classical learning, were reluctant to embrace the new disciplines. University College, London, founded in 1826, provided a broad range of academic studies and offered a course in mechanical philosophy. King's College, London, first taught civil engineering in 1838, and in 1840 Queen Victoria founded the first chair of civil engineering and mechanics at the University of Glasgow, Scot. Rensselaer Polytechnic Institute, founded in 1824, offered the first courses in civil engineering in the United States. The number of universities throughout the world with engineering faculties, including civil engineering, increased rapidly in the 19th and early 20th centuries. Civil engineering today is taught in universities on every continent.

Civil engineering functions. The functions of the civil engineer can be divided into three categories: those performed before construction (feasibility studies, site investigations, and design), those performed during construction (dealing with clients, consulting engineers, and contractors), and those performed after construction (maintenance and research).

Feasibility studies. No major project today is started without an extensive study of the objective and without preliminary studies of possible plans leading to a recommended scheme, perhaps with alternatives. Feasibility studies may cover alternative methods—*e.g.,* bridge versus tunnel, in the case of a water crossing—or, once the method is decided, the choice of route. Both economic and engineering problems must be considered.

Site investigations. A preliminary site investigation is part of the feasibility study, but once a plan has been adopted a more extensive investigation is usually imperative. Money spent in a rigorous study of ground and substructure may save large sums later in remedial works or in changes made necessary in constructional methods.

Soil mechanics

Since the load-bearing qualities and stability of the ground are such important factors in any large-scale construction, it is surprising that a serious study of soil mechanics did not develop until the mid-1930s. Karl von Terzaghi, the chief founder of the science, gives the date of its birth as 1936, when the First International Conference on Soil Mechanics and Foundation Engineering was held at Harvard University and an international society was formed. Today there are specialist societies and journals in many countries, and most universities that have a civil engineering faculty have courses in soil mechanics.

Design. The design of engineering works may require the application of design theory from many fields—*e.g.,* hydraulics, thermodynamics, or nuclear physics. Research in structural analysis and the technology of materials has opened the way for more rational designs, new design concepts, and greater economy of materials. The theory of structures and the study of materials have advanced together as more and more refined stress analysis of structures and systematic testing has been done. Modern designers not only have advanced theories and readily available design data, but structural designs can now be rigorously analyzed by computers.

Construction. The promotion of civil engineering works may be initiated by a private client, but most work is undertaken for large corporations, government authorities, and public boards and authorities. Many of these have their own engineering staffs, but for large specialized projects it is usual to employ consulting engineers.

The role of the consulting engineer

The consulting engineer may be required first to undertake feasibility studies, then to recommend a scheme and quote an approximate cost. The engineer is responsible for the design of the works, supplying specifications, drawings, and legal documents in sufficient detail to seek competitive tender prices. The engineer must compare quotations and recommend acceptance of one of them. Although he is not a party to the contract, the engineer's duties are defined in it; the staff must supervise the construction and the engineer must certify completion of the work. Actions must be consistent with duty to the client; the professional organizations exercise disciplinary control over professional conduct. The consulting engineer's senior representative on the site is the resident engineer.

A phenomenon of recent years has been the turnkey or package contract, in which the contractor undertakes to finance, design, specify, construct, and commission a project in its entirety. In this case, the consulting engineer is engaged by the contractor rather than by the client.

The contractor is usually an incorporated company, which secures the contract on the basis of the consulting engineer's specification and general drawings. The consulting engineer must agree to any variations introduced and must approve the detailed drawings.

Maintenance. The contractor maintains the works to the satisfaction of the consulting engineer. Responsibility for maintenance extends to ancillary and temporary works where these form part of the overall construction. After construction a period of maintenance is undertaken by the contractor, and the payment of the final installment of the contract price is held back until released by the consulting engineer. Central and local government engineering and public works departments are concerned primarily with maintenance, for which they employ direct labour.

Research. Research in the civil engineering field is undertaken by government agencies, industrial foundations, the universities, and other institutions. Most countries have government-controlled agencies, such as the United States Bureau of Standards and the National Physical Laboratory of Great Britain, involved in a broad spectrum of research, and establishments in building research, roads and highways, hydraulic research, water pollution, and other areas. Many are government-aided but depend partly on income from research work promoted by industry.

Branches of civil engineering. In 1828 Thomas Tredgold of England wrote:

> The most important object of Civil Engineering is to improve the means of production and of traffic in states, both for external and internal trade. It is applied in the construction and management of roads, bridges, railroads, aqueducts, canals, river navigation, docks and storehouses, for the convenience of internal intercourse and exchange; and in the construction of ports, harbours, moles, breakwaters and lighthouses; and in the navigation by artificial power for the purposes of commerce.
>
> It is applied to the protection of property where natural powers are the sources of injury, as by embankments for the defence of tracts of country from the encroachments of the sea, or the overflowing of rivers; it also directs the means of applying streams and rivers to use, either as powers to work machines, or as supplies for the use of cities and towns, or for irrigation; as well as the means of removing noxious accumulations, as by the drainage of towns and districts to . . . secure the public health.

Aspects of civil engineering

A modern description would include the production and distribution of energy, the development of aircraft and airports, the construction of chemical process plants and nuclear power stations, and water desalination. These aspects of civil engineering may be considered under the following headings: construction, transportation, maritime and hydraulic engineering, power, and public health.

Construction. Almost all civil engineering contracts include some element of construction work. The development of steel and concrete as building materials had the effect of placing design more in the hands of the civil engineer than the architect. The engineer's analysis of a building problem, based on function and economics, determines the building's structural design.

Transportation. Roman roads and bridges were products of military engineering, but the pavements of McAdam and the bridges of Perronet were the work of the civil engineer. So were the canals of the 18th century and the railways of the 19th, which, by providing bulk transport with speed and economy, lent a powerful impetus to the Industrial Revolution. The civil engineer today is concerned with an even larger transportation field—*e.g.,* traffic studies, design of systems for road, rail, and air, and construction including pavements, embankments, bridges, and tunnels.

Maritime and hydraulic engineering. Harbour construction and shipbuilding are ancient arts. For many developing countries today the establishment of a large, efficient harbour is an early imperative, to serve as the inlet for industrial plant and needed raw materials and the outlet for finished goods. In developed countries the expansion of world trade, the use of larger ships, and the increase in total tonnage call for more rapid and efficient handling. Deeper berths and alongside-handling equipment (for example, for ore) and navigation improvements are the responsibility of the civil engineer.

The development of water supplies was a feature of the earliest civilizations, and the demand for water continues to rise today. In developed countries the demand is for industrial and domestic consumption, but in many parts of the world—*e.g.,* the Indus basin—vast schemes are under construction, mainly for irrigation to help satisfy the food demand, and are often combined with hydroelectric power generation to promote industrial development.

Dam engineering

Dams today are among the largest construction works, and design development is promoted by bodies like the International Commission on Large Dams. The design of large impounding dams in places with population centres close by requires the utmost in safety engineering, with emphasis on soil mechanics and stress analysis. Most governments exercise statutory control of engineers qualified to design and inspect dams.

Power. Civil engineers have always played an important part in mining for coal and metals; the driving of tunnels is a task common to many branches of civil engineering. In the 20th century the design and construction of power stations has advanced with the rapid rise in demand for electric power, and nuclear power stations have added a whole new field of design and construction, involving prestressed concrete pressure vessels for the reactor.

The exploitation of oil fields and the discoveries of natural gas in significant quantities have initiated a radical change in gas production. Shipment in liquid form from the Sahara and piping from the bed of the North Sea have been among the novel developments.

Public health. Drainage and liquid-waste disposal are closely associated with antipollution measures and the reuse of water. The urban development of parts of water catchment areas can alter the nature of runoff, and the training and regulation of rivers produce changes in the pattern of events, resulting in floods and the need for flood prevention and control.

Modern civilization has created problems of solid-waste disposal, from the manufacture of durable goods, such as automobiles and refrigerators, produced in large numbers with a limited life, to the small package, previously disposable, now often indestructible. The civil engineer plays an important role in the preservation of the environment, principally through design of works to enhance rather than to damage or pollute. (J.G.W./Ed.)

MECHANICAL ENGINEERING

Mechanical engineering is the branch of engineering that deals with machines and the production of power. It is particularly concerned with forces and motion.

Steam engine

History. The invention of the steam engine in the latter part of the 18th century, providing a key source of power for the Industrial Revolution, gave an enormous impetus to the development of machinery of all types. As a result, a new major classification of engineering dealing with tools and machines developed, receiving formal recognition in 1847 in the founding of the Institution of Mechanical Engineers in Birmingham, Eng.

Mechanical engineering has evolved from the practice by the mechanic of an art based largely on trial and error to the application by the professional engineer of the scientific method in research, design, and production. The demand for increased efficiency is continually raising the quality of work expected from a mechanical engineer and requiring a higher degree of education and training.

Mechanical engineering functions. Four functions of the mechanical engineer, common to all branches of mechanical engineering, can be cited. The first is the understanding of and dealing with the bases of mechanical science. These include dynamics, concerning the relation between forces and motion, such as in vibration; automatic control; thermodynamics, dealing with the relations among the various forms of heat, energy, and power; fluid flow; heat transfer; lubrication; and properties of materials.

Second is the sequence of research, design, and development. This function attempts to bring about the changes necessary to meet present and future needs. Such work requires a clear understanding of mechanical science, an ability to analyze a complex system into its basic factors, and the originality to synthesize and invent.

Third is production of products and power, which embraces planning, operation, and maintenance. The goal is to produce the maximum value with the minimum investment and cost while maintaining or enhancing longer term viability and reputation of the enterprise or the institution.

Fourth is the coordinating function of the mechanical engineer, including management, consulting, and, in some cases, marketing.

In these functions there is a long continuing trend toward the use of scientific instead of traditional or intuitive methods. Operations research, value engineering, and PABLA (problem analysis by logical approach) are typical titles of such rationalized approaches. Creativity, however, cannot be rationalized. The ability to take the important and unexpected step that opens up new solutions remains in mechanical engineering, as elsewhere, largely a personal and spontaneous characteristic.

Branches of mechanical engineering. *Development of machines for the production of goods.* The high standard of living in the developed countries owes much to mechanical engineering. The mechanical engineer invents machines to produce goods and develops machine tools of increasing accuracy and complexity to build the machines.

The principal lines of development of machinery have been an increase in the speed of operation to obtain high rates of production, improvement in accuracy to obtain quality and economy in the product, and minimization of operating costs. These three requirements have led to the evolution of complex control systems.

The most successful production machinery is that in which the mechanical design of the machine is closely integrated with the control system. A modern transfer (conveyor) line for the manufacture of automobile engines is a good example of the mechanization of a complex series of manufacturing processes. Developments are in hand to automate production machinery further, using computers to store and process the vast amount of data required for manufacturing a variety of components with a small number of versatile machine tools.

Development of machines for the production of power. The steam engine provided the first practical means of generating power from heat to augment the old sources of power from muscle, wind, and water. One of the first challenges to the new profession of mechanical engineering was to increase thermal efficiencies and power; this was done principally by the development of the steam turbine and associated large steam boilers. The 20th century has witnessed a continued rapid growth in the power output of turbines for driving electric generators, together with a steady increase in thermal efficiency and reduction in capital cost per kilowatt of large power stations. Finally, mechanical engineers acquired the resource of nuclear energy, whose application has demanded an exceptional standard of reliability and safety involving the solution of entirely new problems (see *Nuclear engineering* below).

Complex control systems

The mechanical engineer is also responsible for the much smaller internal combustion engines, both reciprocating (gasoline and diesel) and rotary (gas-turbine and Wankel) engines, with their widespread transport applications. In the transportation field generally, in air and space as well as on land and sea, the mechanical engineer has created the equipment and the power plant, collaborating increasingly with the electrical engineer, especially in the development of suitable control systems.

Development of military weapons. The skills applied to war by the mechanical engineer are similar to those required in civilian applications, though the purpose is to enhance destructive power rather than to raise creative efficiency. The demands of war have channeled huge resources into technical fields, however, and led to developments that have profound benefits in peace. Jet aircraft and nuclear reactors are notable examples.

Environmental control. The earliest efforts of mechanical engineers were aimed at controlling the human environment by draining and irrigating land and by ventilating mines. Refrigeration and air conditioning are examples of the use of modern mechanical devices to control the environment.

Many of the products of mechanical engineering, together with technological developments in other fields, give rise to noise, the pollution of water and air, and the dereliction of land and scenery. The rate of production, both of goods and power, is rising so rapidly that regeneration by natural forces can no longer keep pace. A rapidly growing field for mechanical engineers and others is environmental control, comprising the development of machines and processes that will produce fewer pollutants and of new equipment and techniques that can reduce or remove the pollution already generated. (J.F.Br./P.McG.R./Ed.)

Side effects of development

CHEMICAL ENGINEERING

Chemical engineering is the development of processes and the design and operation of plants in which materials undergo changes in physical or chemical state on a technical scale. Applied throughout the process industries, it is founded on the principles of chemistry, physics, and mathematics. The laws of physical chemistry and physics govern the practicability and efficiency of chemical engineering operations. Energy changes, deriving from thermodynamic considerations, are particularly important. Mathematics is a basic tool in optimization and modeling. Optimization means arranging materials, facilities, and energy to yield as productive and economical an operation as possible. Modeling is the construction of theoretical mathematical prototypes of complex process systems, commonly with the aid of computers.

History. Chemical engineering is as old as the process industries. Its heritage dates from the fermentation and evaporation processes operated by early civilizations. Modern chemical engineering emerged with the development of large-scale, chemical-manufacturing operations in the second half of the 19th century. Throughout its development as an independent discipline, chemical engineering has been directed toward solving problems of designing and operating large plants for continuous production.

Manufacture of chemicals in the mid-19th century consisted of modest craft operations. Increase in demand, public concern at the emission of noxious effluents, and competition between rival processes provided the incentives for greater efficiency. This led to the emergence of combines with resources for larger operations and caused the transition from a craft to a science-based industry. The result was a demand for chemists with knowledge of manufacturing processes, known as industrial chemists or chemical technologists. The term chemical engineer was in general use by about 1900. Despite its emergence in traditional chemicals manufacturing, it was through its role in the development of the petroleum industry that chemical engineering became firmly established as a unique discipline. The demand for plants capable of operating physical separation processes continuously at high levels of efficiency was a challenge that could not be met by the traditional chemist or mechanical engineer.

A landmark in the development of chemical engineering was the publication in 1901 of the first textbook on the subject, by George E. Davis, a British chemical consultant. This concentrated on the design of plant items for specific operations. The notion of a processing plant encompassing a number of operations, such as mixing, evaporation, and filtration, and of these operations being essentially similar, whatever the product, led to the concept of unit operations. This was first enunciated by the American chemical engineer Arthur D. Little in 1915 and formed the basis for a classification of chemical engineering that dominated the subject for the next 40 years. The number of unit operations—the building blocks of a chemical plant—is not large. The complexity arises from the variety of conditions under which the unit operations are conducted.

First chemical engineering textbook

In the same way that a complex plant can be divided into basic unit operations, so chemical reactions involved in the process industries can be classified into certain groups, or unit processes (*e.g.*, polymerizations, esterifications, and

nitrations), having common characteristics. This classification into unit processes brought rationalization to the study of process engineering.

The unit approach suffered from the disadvantage inherent in such classifications: a restricted outlook based on existing practice. Since World War II, closer examination of the fundamental phenomena involved in the various unit operations has shown these to depend on the basic laws of mass transfer, heat transfer, and fluid flow. This has given unity to the diverse unit operations and has led to the development of chemical engineering science in its own right; as a result, many applications have been found in fields outside the traditional chemical industry.

Study of the fundamental phenomena upon which chemical engineering is based has necessitated their description in mathematical form and has led to more sophisticated mathematical techniques. The advent of digital computers has allowed laborious design calculations to be performed rapidly, opening the way to accurate optimization of industrial processes. Variations due to different parameters, such as energy source used, plant layout, and environmental factors, can be predicted accurately and quickly so that the best combination can be chosen.

Chemical engineering functions. Chemical engineers are employed in the design and development of both processes and plant items. In each case, data and predictions often have to be obtained or confirmed with pilot experiments. Plant operation and control is increasingly the sphere of the chemical engineer rather than the chemist. Chemical engineering provides an ideal background for the economic evaluation of new projects and, in the plant construction sector, for marketing.

Branches of chemical engineering. The fundamental principles of chemical engineering underlie the operation of processes extending well beyond the boundaries of the chemical industry, and chemical engineers are employed in a range of operations outside traditional areas. Plastics, polymers, and synthetic fibres involve chemical-reaction engineering problems in their manufacture, with fluid flow and heat transfer considerations dominating their fabrication. The dyeing of a fibre is a mass-transfer problem. Pulp and paper manufacture involve considerations of fluid flow and heat transfer. While the scale and materials are different, these again are found in modern continuous production of foodstuffs. The pharmaceuticals industry presents chemical engineering problems, the solutions of which have been essential to the availability of modern drugs. The nuclear industry makes similar demands on the chemical engineer, particularly for fuel manufacture and reprocessing. Chemical engineers are involved in many sectors of the metals processing industry, which extends from steel manufacture to separation of rare metals.

Non-traditional employment

Further applications of chemical engineering are found in the fuel industries. In the second half of the 20th century, considerable numbers of chemical engineers have been involved in space exploration, from the design of fuel cells to the manufacture of propellants. Looking to the future, it is probable that chemical engineering will provide the solution to at least two of the world's major problems: supply of adequate fresh water in all regions through desalination of seawater and environmental control through prevention of pollution. (C.Ha./Ed.)

ELECTRICAL AND ELECTRONICS ENGINEERING

Electrical engineering deals with the practical applications of electricity in all its forms, including those of the field of electronics. Electronics engineering is that branch of electrical engineering concerned with the uses of the electromagnetic spectrum and with the application of such electronic devices as integrated circuits, transistors, and vacuum tubes. In engineering practice, the distinction between electrical engineering and electronics is based on the comparative strength of the electric currents used. In this sense, electrical engineering is the branch dealing with "heavy current"—that is, electric light and power systems and apparatuses—whereas electronics engineering deals with such "light current" applications as wire and radio communication, the stored-program electronic computer, radar, and automatic control systems.

The distinction between the fields has become less sharp with recent technical progress. For example, in the high-voltage transmission of electric power, large arrays of electronic devices are used to convert transmission-line current at power levels in the tens of megawatts. Moreover, in the regulation and control of interconnected power systems, electronic computers are used to compute requirements much more rapidly and accurately than is possible by manual methods. (D.G.F.)

History. Electrical phenomena attracted the attention of European thinkers as early as the 17th century. Beginning as a mathematically oriented science, the field has remained primarily in that form; mathematical predication often precedes laboratory demonstration. The most noteworthy pioneers include Ludwig Wilhelm Gilbert and Georg Simon Ohm of Germany, Hans Christian Ørsted of Denmark, André-Marie Ampère of France, Alessandro Volta of Italy, Joseph Henry of the United States, and Michael Faraday of England. Electrical engineering may be said to have emerged as a discipline in 1864 when the Scottish physicist James Clerk Maxwell summarized the basic laws of electricity in mathematical form and predicted that radiation of electromagnetic energy would occur in a form that later became known as radio waves. In 1887 the German physicist Heinrich Hertz experimentally demonstrated the existence of radio waves.

Early theorists

The first practical application of electricity was the telegraph, invented by Samuel F.B. Morse in 1837. The need for electrical engineers was not felt until some 40 years later, upon the invention of the telephone (1876) by Alexander Graham Bell and of the incandescent lamp (1878) by Thomas A. Edison. These devices and Edison's first central generating plant in New York City (1882) created a large demand for men trained to work with electricity.

First practical application

The discovery of the "Edison effect," a flow of current through the vacuum of one of his lamps, was the first observation of current in space. Hendrick Antoon Lorentz of The Netherlands predicted the electron theory of electrical charge in 1895, and in 1897 J.J. Thomson of England showed that the Edison effect current was indeed caused by negatively charged particles (electrons). This led to the work of Guglielmo Marconi of Italy, Lee De Forest of the United States, and many others, which laid the foundations of radio engineering. In 1930 the term electronics was introduced to embrace radio and the industrial applications of electron tubes. Since 1947, when the transistor was invented by John Bardeen, William H. Brattain, and William B. Shockley, electronics engineering has been dominated by the applications of such solid-state electronic devices as the transistor, the semiconductor diode, and the integrated circuit. (J.D.R./D.G.F./Ed.)

Electrical and electronics engineering functions. *Research.* The functions performed by electrical and electronics engineers include (1) basic research in physics, other sciences, and applied mathematics in order to extend knowledge applicable to the field of electronics, (2) applied research based on the findings of basic research and directed at discovering new applications and principles of operation, (3) development of new materials, devices, assemblies, and systems suitable for existing or proposed product lines, (4) design of devices, equipment, and systems for manufacture, (5) field-testing of equipment and systems, (6) establishment of quality control standards to be observed in manufacture, (7) supervision of manufacture and production testing, (8) postproduction assessment of performance, maintenance, and repair, and (9) engineering management, or the direction of research, development, engineering, manufacture, and marketing and sales.

Consulting. The rapid proliferation of new discoveries, products, and markets in the electrical and electronics industries has made it difficult for workers in the field to maintain the range of skills required to manage their activities. Consulting engineers, specializing in new fields, are employed to study and recommend courses of action.

The educational background required for these functions tends to be highest in basic and applied research. In most major laboratories a doctorate in science or engineering is

required to fill leadership roles. Most positions in design, product development, and supervision of manufacture and quality control require a master's degree. In the high-technology industries typical of modern electronics, an engineering background at not less than the bachelor's level is required to assess competitive factors in sales engineering to guide marketing strategy.

Branches of electrical and electronics engineering. The largest of the specialized branches of electrical engineering, the branch concerned with the electronic computer, was introduced during World War II. The field of computer science and engineering has attracted members of several disciplines outside electronics, notably logicians, linguists, and applied mathematicians.

Another very large field is that concerned with electric light and power and their applications. Specialities within the field include the design, manufacture, and use of turbines, generators, transmission lines, transformers, motors, lighting systems, and appliances.

A third major field is that of communications, which comprises not only telegraphy and telephony but also satellite communications and the transmission of voice and data by laser signals through optical-fibre networks. The communication of digital data among computers connected by wire, microwave, and satellite circuits is now a major enterprise that has built a strong bond between computer and communications specialists.

Applications for other fields

The applications of electricity and electronics to other fields of science have expanded since World War II. Among the sciences represented are medicine, biology, oceanography, geoscience, nuclear science, laser physics, sonics and ultrasonics, and acoustics. Theoretical specialties within electronics include circuit theory, information theory, radio-wave propagation, and microwave theory.

Another important speciality concerns improvements in materials and components used in electrical and electronics engineering, such as conductive, magnetic, and insulating materials and the semiconductors used in solid-state devices. One of the most active areas is the development of new electronic devices, particularly the integrated circuits used in computers and other digital systems.

The development of electronic systems—equipment for consumers, such as radios, television sets, stereo equipment, video games, and home computers—occupies a large number of engineers. Another field is the application of computers and radio systems to automobiles, ships, and other vehicles. The field of aerospace electronic systems includes navigation aids for aircraft, automatic pilots, altimeters, and radar for traffic control, blind landing, and collision prevention. Many of these devices are also widely used in the marine services. (D.G.F./Ed.)

PETROLEUM ENGINEERING

Petroleum engineering is a specialized engineering discipline whose origins lie in both mining engineering and geology. The petroleum engineer, whose aim is to extract gaseous and liquid hydrocarbon products from the earth, is concerned with drilling, producing, processing, and transporting these products and handling all the related economic and regulatory considerations.

History. The foundations of petroleum engineering were established during the 1890s in California. There geologists were employed to correlate oil-producing zones and water zones from well to well to prevent extraneous water from entering oil-producing zones. From this came the recognition of the potential for applying technology to oil-field development. The American Institute of Mining and Metallurgical Engineers (AIME) established a Technical Committee on Petroleum in 1914. In 1957 the name of the AIME was changed to the American Institute of Mining, Metallurgical, and Petroleum Engineers.

Introduction of petroleum technology courses

Petroleum technology courses were introduced at the University of Pittsburgh, Pa., in 1910 and included courses in oil and gas law and industry practices; in 1915 the university granted the first degree in petroleum engineering. Also in 1910 the University of California at Berkeley offered its first courses in petroleum engineering and in 1915 established a four-year curriculum in petroleum engineering. After these pioneering efforts, professional programs spread throughout the United States and other countries.

From 1900 to 1920 petroleum engineering focused on drilling problems, such as establishing casing points for water shutoff, designing casing strings, and improving the mechanical operations in drilling and well pumping. In the 1920s petroleum engineers sought means to improve drilling practices and to improve well design by use of proper tubing sizes, chokes, and packers. They designed new forms of artificial lift, primarily rod pumping and gas lift, and studied the ways in which methods of production affected gas–oil ratios and rates of production. The technology of drilling fluids was advanced, and directional drilling became a common practice.

The economic crisis that resulted from abundant discoveries in about 1930, notably in the giant East Texas Field, caused petroleum engineering to focus on the entire oil–water–gas reservoir system rather than on the individual well. Studying the optimum spacing of wells in an entire field led to the concept of reservoir engineering. During this period the mechanics of drilling and production were not neglected. Drilling penetration rates increased approximately 100 percent from 1932 to 1937.

Petrophysics (determination of fluid and rock characteristics) was introduced late in the 1930s. By 1940 electric logging had developed to the state that estimates could be made of oil and water saturations in the reservoir rocks.

After World War II, petroleum engineers continued to refine the techniques of reservoir analysis and petrophysics. The outstanding event of the 1950s was development of the offshore oil industry and a whole new technology. At first little was known of such matters as wave heights and wave forces. The oceanographer and marine engineer thus joined with the petroleum engineer to initiate design standards. Shallow-water drilling barges evolved into mobile platforms, then into jack-up barges, and finally into semi-submersible and floating drilling ships.

Specialization

Branches of petroleum engineering. During the evolution of petroleum engineering, the areas of specialization developed: drilling engineering, production engineering, reservoir engineering, and petrophysical engineering. In each specialization engineers from other disciplines (mechanical, civil, electrical, geological, chemical) freely entered, and their contributions were significant; however, it remained the unique role of the petroleum engineer to integrate all the specializations into an efficient system of oil and gas drilling, production, and processing.

Drilling engineering was among the first applications of technology to oil-field practices. The drilling engineer is responsible for the design of the earth-penetration techniques, the selection of casing and safety equipment, and, often, the direction of the operations. These functions involve understanding the nature of the rocks to be penetrated, the stresses in these rocks, and the techniques available to drill into and control the underground reservoirs. Because modern drilling involves organizing a vast array of machinery and materials, investing huge funds, and acknowledging the safety and welfare of the general public, the engineer must develop the skills of supervision, management, and negotiation.

The production engineer's work begins upon completion of the well—directing the selection of producing intervals and making arrangements for various accessories, controls, and equipment. Later his work involves controlling and measuring the produced fluids (oil, gas, and water), designing and installing gathering and storage systems, and delivering the raw products (gas and oil) to pipeline companies and other transportation agents. He is also involved in such matters as corrosion prevention, well performance, and formation treatments to stimulate production. As in all branches of petroleum engineering, the production engineer cannot view the in-hole or surface processing problems in isolation but must fit solutions into the complete reservoir, well, and surface system.

Reservoir engineers are concerned with the physics of oil and gas distribution and their flow through porous rocks—the various hydrodynamic, thermodynamic, gravitational, and other forces involved in the rock–fluid system. They are responsible for analyzing the rock–fluid system, establishing efficient well-drainage patterns, forecasting the

performance of the oil or gas reservoir, and introducing methods for maximum efficient production.

To understand the reservoir rock–fluid system, the drilling, production, and reservoir engineers draw assistance from the petrophysical, or formation-evaluation, engineer, who provides tools and analytical techniques for determining rock and fluid characteristics. The petrophysical engineer measures the acoustic, radioactive, and electrical properties of the rock–fluid system and takes samples of the rocks and well fluids to determine porosity, permeability, and fluid content in the reservoir.

(B.D.H./Ed.)

AEROSPACE ENGINEERING

Aerospace engineering is the study of the design, development, and operation of vehicles operating in the Earth's atmosphere or in outer space. In 1958 the first definition of aerospace engineering appeared, considering the Earth's atmosphere and the space above it as a single realm for development of flight vehicles. Today the more encompassing aerospace definition has commonly replaced the terms aeronautical engineering and astronautical engineering.

The design of a flight vehicle demands a knowledge of many engineering disciplines. It is rare that one person takes on the entire task; instead, most companies have design teams specialized in the sciences of aerodynamics, propulsion systems, structural design, materials, avionics, and stability and control systems. No single design can optimize all of these sciences, but rather there exist compromised designs that incorporate the vehicle specifications, available technology, and economic feasibility.

History. *Aeronautical engineering.* The roots of aeronautical engineering can be traced to the early days of mechanical engineering, to inventors' concepts, and to the initial studies of aerodynamics, a branch of theoretical physics. The earliest sketches of flight vehicles were drawn by Leonardo da Vinci, who suggested two ideas for sustentation. The first was an ornithopter, a flying machine using flapping wings to imitate the flight of birds. The second idea was an aerial screw, the predecessor of the helicopter. Manned flight was first achieved in 1783, in a hot-air balloon designed by the French brothers Joseph-Michel and Jacques-Étienne Montgolfier. Aerodynamics became a factor in balloon flight when a propulsion system was considered for forward movement. Benjamin Franklin was one of the first to propose such an idea, which led to the development of the dirigible. The power-driven balloon was invented by Henri Gifford, a Frenchman, in 1852. The invention of lighter-than-air vehicles occurred independently of the development of aircraft. The breakthrough in aircraft development came in 1799 when Sir George Cayley, an English baron, drew an airplane incorporating a fixed wing for lift, an empennage (consisting of horizontal and vertical tail surfaces for stability and control), and a separate propulsion system. Because engine development was virtually nonexistent, Cayley turned to gliders, building the first successful one in 1849. Gliding flights established a data base for aerodynamics and aircraft design. Otto Lilienthal, a German scientist, recorded more than 2,000 glides in a five-year period, beginning in 1891. Lilienthal's work was followed by the American aeronaut Octave Chanute, a friend of the American brothers Orville and Wilbur Wright, the fathers of modern manned flight.

The first manned flight

Following the first sustained flight of a heavier-than-air vehicle in 1903, the Wright brothers refined their design, eventually selling airplanes to the U.S. Army. The first major impetus to aircraft development occurred during World War I, when aircraft were designed and constructed for specific military missions, including fighter attack, bombing, and reconnaissance. The end of the war marked the decline of military high-technology aircraft and the rise of civil air transportation. Many advances in the civil sector were due to technologies gained in developing military and racing aircraft. A successful military design that found many civil applications was the U.S. Navy Curtiss NC-4 flying boat, powered by four 400-horsepower V-12 Liberty engines. It was the British, however, who paved the way in civil aviation in 1920 with a 12-passenger Handley-Page transport. Aviation boomed after Charles A. Lindbergh's solo flight across the Atlantic Ocean in 1927. Advances in metallurgy led to improved strength-to-weight ratios and, coupled with a monocoque design, enabled aircraft to fly farther and faster. Hugo Junkers, a German, built the first all-metal monoplane in 1910, but the design was not accepted until 1933, when the Boeing 247-D entered service. The twin-engine design of the latter established the foundation of modern air transport.

The advent of the turbine-powered airplane dramatically changed the air transportation industry. Germany and Britain were concurrently developing the jet engine, but it was a German Heinkel He 178 that made the first jet flight on Aug. 27, 1939. Even though World War II accelerated the growth of the airplane, the jet aircraft was not introduced into service until 1944, when the British Gloster Meteor became operational, shortly followed by the German Me 262. The first practical American jet was the Lockheed F-80, which entered service in 1945.

Commercial aircraft after World War II continued to use the more economical propeller method of propulsion. The efficiency of the jet engine was increased, and in 1949 the British de Havilland Comet inaugurated commercial jet transport flight. The Comet, however, experienced structural failures that curtailed the service, and it was not until 1958 that the highly successful Boeing 707 jet transport began nonstop transatlantic flights. While civil aircraft designs utilize most new technological advancements, the transport and general aviation configurations have changed only slightly since 1960. Because of escalating fuel and hardware prices, the development of civil aircraft has been dominated by the need for economical operation.

Technological improvements in propulsion, materials, avionics, and stability and controls have enabled aircraft to grow in size, carrying more cargo faster and over longer distances. While aircraft are becoming safer and more efficient, they are also now very complex. Today's commercial aircraft are among the most sophisticated engineering achievements of the day.

Growing complexity of aircraft

Smaller, more fuel-efficient airliners are being developed. The use of turbine engines in light general aviation and commuter aircraft is being explored, along with more efficient propulsion systems, such as the propfan concept. Using satellite communication signals, onboard microcomputers can provide more accurate vehicle navigation and collision-avoidance systems. Digital electronics coupled with servo mechanisms can increase efficiency by providing active stability augmentation of control systems. New composite materials providing greater weight reduction; inexpensive one-man, lightweight, noncertified aircraft, referred to as ultralights; and alternate fuels such as ethanol, methanol, synthetic fuel from shale deposits and coal, and liquid hydrogen are all being explored. Aircraft designed for vertical and short takeoff and landing, which can land on runways one-tenth the normal length, are being developed. Hybrid vehicles such as the Bell XV-15 tilt-rotor already combine the vertical and hover capabilities of the helicopter with the speed and efficiency of the airplane. Although environmental restrictions and high operating costs have limited the success of the supersonic civil transport, the appeal of reduced traveling time justifies the examination of a second generation of supersonic aircraft.

Aerospace engineering. The use of rocket engines for aircraft propulsion opened a new realm of flight to the aeronautical engineer. Robert H. Goddard, an American, developed, built, and flew the first successful liquid-propellant rocket on March 16, 1926. Goddard proved that flight was possible at speeds greater than the speed of sound and that rockets can work in a vacuum. The major impetus in rocket development came in 1938 when the American James Hart Wyld designed, built, and tested the first U.S. regeneratively cooled liquid rocket engine. In 1947 Wyld's rocket engine powered the first supersonic research aircraft, the Bell X-1, flown by the U.S. Air Force captain Charles E. Yeager. Supersonic flight offered the aeronautical engineer new challenges in propulsion, structures and materials, high-speed aeroelasticity, and transonic, supersonic, and hypersonic aerodynamics. The

experience gained in the X-1 tests led to the development of the X-15 research rocket plane, which flew more than 700 flights over a 22-year period. The X-15 established an extensive data base in transonic and supersonic flight (up to five times the speed of sound) and revealed vital information concerning the upper atmosphere.

The space exploration race

The late 1950s and '60s marked a period of intense growth for astronautical engineering. In 1957 the U.S.S.R. orbited Sputnik I, the world's first artificial satellite, which triggered a space exploration race with the United States. In 1961 the U.S. president John F. Kennedy recommended to Congress to undertake the challenge of "landing a man on the moon and returning him safely to the earth" by the end of the 1960s. This commitment was fulfilled on July 20, 1969, when astronauts Neil A. Armstrong and Edwin E. Aldrin, Jr., landed on the Moon.

The 1970s began the decline of the U.S. manned spaceflights. The exploration of the Moon was replaced by unmanned voyages to Jupiter, Saturn, and other planets. Eventually the exploitation of space was redirected from conquering distant planets to providing a better understanding of the human environment. Artificial satellites provide statistical data pertaining to geographic formations, oceanic and atmospheric movements, radiation mapping, and worldwide communications. The frequency of U.S. spaceflights in the 1960s and '70s led to the development of a reusable, low-orbital-altitude space shuttle. With several successful flights in the 1980s, the shuttle inaugurated a new age of commercially viable space vehicles.

Aerospace engineering functions. In most countries, governments are the aerospace industry's largest customers, and most engineers work on the design of military vehicles. The largest demand for aerospace engineers comes from the transport and fighter aircraft, missile, spacecraft, and general aviation industries. The typical aerospace engineer holds a bachelor's degree, but there are many engineers holding master's or doctorate degrees (or their equivalents) in various disciplines associated with aerospace-vehicle design, development, and testing.

The U.S. National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) are governmental organizations that employ many engineers for research, development, testing, and procurement of military vehicles. Government agencies award and monitor industrial contracts ranging from engineering problem studies to design and fabrication of hardware. Universities receive limited funding, primarily for analytical research. Some of the larger institutions, however, are developing or expanding flight-research facilities and increasing faculty members in an effort to increase productivity in both research and testing.

The design of a flight vehicle is a complex and time-consuming procedure requiring the integration of many engineering technologies. Supporting teams are formed to provide expertise in these technologies, resulting in a completed design that is the best compromise of all the engineering disciplines. Usually the support teams are supervised by a project engineer or chief designer for technical guidance and by a program manager responsible for program budgets and schedules. Because of the ever-increasing requirement for advanced technology and the high cost and high risk associated with complex flight vehicles, many research and development programs are canceled before completion.

The phases of the design process

The design process can be dissected into five phases and is the same for most aerospace products. Phase one is a marketing analysis to determine customer specifications or requirements. Aerospace engineers are employed to examine technical, operational, or financial problems. The customer's requirements are established and then passed on to the conceptual design team for the second phase.

The conceptual design team generally consists of aerospace engineers, who make the first sketch attempt to determine the vehicle's size and configuration. Preliminary estimates of the vehicle's performance, weight, and propulsion systems are made. Performance parameters include range, speed, drag, power required, payload, and takeoff and landing distances. Parametric trade studies are conducted to optimize the design, but configuration details usually change. This phase may take from a few months to years for major projects.

Phase three is the preliminary design phase. The optimized vehicle design from phase two is used as the starting point. Aerospace engineers perform computer analyses on the configuration; then wind-tunnel models are built and tested. Flight control engineers study dynamic stability and control problems. Propulsion groups supply data necessary for engine selection. Interactions between the engine inlet and vehicle frame are studied. Civil, mechanical, and aerospace engineers analyze the bending loads, stresses, and deflections on the wing, airframe, and other components. Material science engineers aid in selecting low-weight, high-strength materials and may conduct aeroelastic and fatigue tests. Weight engineers make detailed estimates of individual component weights. As certain parameters drive the vehicle design, the preliminary designers are often in close contact with both the conceptual designers and the marketing analysts. The time involved in the preliminary design phase depends on the complexity of the problem but usually takes from six to 24 months.

Phase four, the detailed design phase, involves construction of a prototype. Mechanical engineers, technicians, and draftsmen help lay out the drawings necessary to construct each component. Full-scale mock-ups are built of cardboard, wood, or other inexpensive materials to aid in the subsystem layout. Subsystem components are built and bench-tested, and additional wind-tunnel testing is performed. This phase takes from one to three years.

The final phase concerns flight-testing the prototype. Engineers and test pilots work together to assure that the vehicle is safe and performs as expected. If the prototype is a commercial transport aircraft, the vehicle must meet the requirements specified by government organizations such as the Federal Aviation Administration in the United States and the Civil Aviation Authority in the United Kingdom. Prototype testing is usually completed in one year but can take much longer because of unforeseen contingencies. The time required from the perception of a customer's needs to delivery of the product can be as long as 10 to 15 years depending on the complexity of the design, the political climate, and the availability of funding.

High-speed computers have now enabled complex aerospace engineering problems to be analyzed rapidly. More extensive computer programs, many written by aerospace engineers, are being formulated to aid the engineer in designing new configurations.

Branches of aerospace engineering. The aerospace engineer is armed with an extensive background suitable for employment in most positions traditionally occupied by mechanical engineers as well as limited positions in the other various engineering disciplines. The transportation, construction, communication, and energy industries provide the most opportunities for non-aerospace applications.

Non-aerospace applications

Because land and sea vehicles are designed for optimum speed and efficiency, the aerospace engineer has become a prominent member of the design teams. Because up to half of the power required to propel a vehicle is due to the resistance of the air, the configuration design of low-drag automobiles, trains, and boats offers better speed and fuel economy. The presence of the aerospace engineer in the automobile industry is evident from the streamlined shapes of cars and trucks that evolved during the late 20th century, at a time when gasoline prices were escalating and the aerospace industry was in a lull. Airline companies employ engineers as performance analysts, crash investigators, and consultants. The Federal Aviation Administration and other governmental organizations use the technical expertise of the aerospace engineer in various capacities.

The construction of large towers, buildings, and bridges requires predictions of aerodynamic forces and the creation of an optimum design to minimize these forces. The consideration of aerodynamic forces of flat surfaces such as the side of a building or superstructure is not new. In 1910 Alexandre-Gustave Eiffel achieved remarkable experimental results measuring the wind resistance of a flat plate, using the Eiffel Tower as a test platform.

Many companies benefit not from the advanced hard-

ware developments of aerospace technology but by the understanding and application of aerospace methodology. Companies engaged in satellite communications require an understanding of orbital mechanics, trajectories, acceleration forces, and aerodynamic heating and an overall knowledge of the spacecraft industry. Advanced aerodynamic design of airfoils and rotor systems is applied in an effort to improve the efficiency of propellers, windmills, and turbine engines. The impact of aerospace technology has trickled down to many companies engaged in the research and development of flight simulation, automatic controls, materials, dynamics, robotics, medicine, and other high-technology fields. (K.A.St./Ed.)

BIOENGINEERING

Bioengineering is the application of engineering knowledge to the fields of medicine and biology. The bioengineer must be well grounded in biology and have engineering knowledge that is broad, drawing upon electrical, chemical, mechanical, and other engineering disciplines. The bioengineer may work in any of a large range of areas. One of these is the provision of artificial means to assist defective body functions—such as hearing aids, artificial limbs, and supportive or substitute organs. In another direction, the bioengineer may use engineering methods to achieve biosynthesis of animal or plant products—such as for fermentation processes.

History. Before World War II the field of bioengineering was essentially unknown, and little communication or interaction existed between the engineer and the life scientist. A few exceptions, however, should be noted. The agricultural engineer and the chemical engineer, involved in fermentation processes, have always been bioengineers in the broadest sense of the definition since they deal with biological systems and work with biologists. The civil engineer, specializing in sanitation, has applied biological principles in the work. Mechanical engineers have worked with the medical profession for many years in the development of artificial limbs. Another area of mechanical engineering that falls in the field of bioengineering is the air-conditioning field. In the early 1920s engineers and physiologists were employed by the American Society of Heating and Ventilating Engineers to study the effects of temperature and humidity on humans and to provide design criteria for heating and air-conditioning systems.

Today there are many more examples of interaction between biology and engineering, particularly in the medical and life-support fields. In addition to an increased awareness of the need for communication between the engineer and the associate in the life sciences, there is an increasing recognition of the role the engineer can play in several of the biological fields, including human medicine, and, likewise, an awareness of the contributions biological science can make toward the solution of engineering problems.

Much of the increase in bioengineering activity can be credited to electrical engineers. In the 1950s bioengineering meetings were dominated by sessions devoted to medical electronics. Medical instrumentation and medical electronics continue to be major areas of interest, but biological modeling, blood-flow dynamics, prosthetics, biomechanics (dynamics of body motion and strength of materials), biological heat transfer, biomaterials, and other areas are now included in conference programs.

The merging of medical and engineering needs

Bioengineering developed out of specific desires or needs: the desire of surgeons to bypass the heart, the need for replacement organs, the requirement for life support in space, and many more. In most cases the early interaction and education were a result of personal contacts between physician, or physiologist, and engineer. Communication between the engineer and the life scientist was immediately recognized as a problem. Most engineers who wandered into the field in its early days probably had an exposure to biology through a high-school course and no further work. To overcome this problem, engineers began to study not only the subject matter but also the methods and techniques of their counterparts in medicine, physiology, psychology, and biology. Much of the information was self-taught or obtained through personal association and discussions. Finally, recognizing a need to assist in overcoming the communication barrier as well as to prepare engineers for the future, engineering schools developed courses and curricula in bioengineering.

Branches of bioengineering. *Medical engineering.* Medical engineering concerns the application of engineering principles to medical problems, including the replacement of damaged organs, instrumentation, and the systems of health care, including diagnostic applications of computers.

Agricultural engineering. This includes the application of engineering principles to the problems of biological production and to the external operations and environment that influence this production.

Bionics. Bionics is the study of living systems so that the knowledge gained can be applied to the design of physical systems.

Biochemical engineering. Biochemical engineering includes fermentation engineering, application of engineering principles to microscopic biological systems that are used to create new products by synthesis, including the production of protein from suitable raw materials.

Human-factors engineering. This concerns the application of engineering, physiology, and psychology to the optimization of the human–machine relationship.

Environmental health engineering. Also called bioenvironmental engineering, this field concerns the application of engineering principles to the control of the environment for the health, comfort, and safety of human beings. It includes the field of life-support systems for the exploration of outer space and the ocean. (Ed.)

NUCLEAR ENGINEERING

Nuclear engineering is concerned with the control and use of energy and radiation released from nuclear reactions. It encompasses the development, design, and construction of power reactors, naval-propulsion reactors, nuclear fuel-cycle facilities, and radioactive-waste disposal facilities; the development and production of nuclear weapons; and the production and application of radioisotopes.

History. Nuclear engineering began with the first major demonstrations of the utilization of nuclear energy: the development of nuclear weapons and nuclear reactors.

The Manhattan Project

The World War II Manhattan Project, under which the U.S. government built, in a relatively short period, such facilities as production reactors, chemical-reprocessing plants, test and research reactors, and weapons production facilities, stands out as a monumental engineering feat. Engineers in early programs had to learn about a host of nuclear-related subjects, ranging from reactor theory and reactor control to radioactivity and the behaviour of material under irradiation. They were educated on the job by nuclear scientists and physicists, first through personal discussions and later through seminars and classes. Many of those who entered the field had been educated in other engineering disciplines—mechanical, electrical, chemical, and so on. Nuclear engineering continues today to be a strongly interdisciplinary activity.

Early schools. In the late 1940s, as the many potential peaceful uses of nuclear energy became evident, two schools of reactor technology were established, one in Tennessee at Oak Ridge National Laboratory and another in Illinois at Argonne National Laboratory.

In 1946 Clinch College was established at Oak Ridge. In its first year 35 American participants from universities, industry, the U.S. Navy, and government agencies took courses in nuclear technology. They attended lectures, conducted laboratory experiments, and gained hands-on experience in operating nuclear reactors.

In 1950 Clinch College was succeeded by the Oak Ridge School of Reactor Technology (ORSORT). The participants were again selected from academic, government, and industry sectors. In addition to lectures and laboratory work, the students were assigned to teams working on the development of new concepts. Several concepts developed by these teams later grew into major research and development programs, including the high-flux isotope reactor, the molten-salt reactor, and several nuclear propulsion schemes. ORSORT was disbanded in 1965 because nuclear engineering programs had by that time become widely available at universities and colleges.

The International School of Nuclear Science and Engineering was established at Argonne National Laboratory in 1955. The school was created to meet the international need for trained scientists and engineers, and its program was conducted jointly by Argonne National Laboratory, North Carolina State College, and Pennsylvania State University. Basic course work was presented at the universities in a 17-week program combining lecture with laboratory experience. More advanced work, including lectures and participation in design and laboratory projects, was given in a second 17-week program at the International School at Argonne. In 1960 the basic course work was discontinued, and the program was redirected to serve more advanced and experienced students from abroad. In recognition of the worldwide growth of programs and facilities to provide basic nuclear training at universities and laboratories, the program at Argonne was discontinued in 1964.

University programs. In 1950 the first full-fledged nuclear engineering curriculum offered for college credit was established at North Carolina State College. By 1952 several schools had graduate programs in nuclear engineering. Most of these programs consisted of two or three courses, providing a background on reactor physics, reactor control, heat transfer, radiation effects, and shielding.

With the support of the U.S. Atomic Energy Commission's Division of Nuclear Education and Training, the curricula and the number of schools in the United States continued to increase. By 1965, 61 schools were offering nuclear engineering programs. The programs had grown in diverse directions, however, and it became apparent that it was desirable to develop a consensus among educators about nuclear engineering education. To meet this need, a joint committee of the American Nuclear Society and the American Society of Engineering Education developed basic educational criteria. The committee members came from industry, national laboratories, and universities with nuclear engineering programs. The committee's "Report on Objective Criteria in Nuclear Engineering Education" had a major influence in shaping nuclear engineering curricula around the world and did much to establish nuclear engineering as a distinct discipline.

Education criteria established

Nuclear engineering functions. *Research and development.* Research and development entails the conception and development of new materials, processes, components, and systems for nuclear facilities and the development of analytical methods and experimental procedures for use in the development, analysis, design, and control of fission and fusion systems.

Design. Another area of emphasis is the engineering design of such items as fuel elements, reactor-core supports, reflectors, thermal shields, biological shields, instrumentation and control systems, and safety systems.

Fuel management. Fuel management involves specifying, procuring, and managing fuel throughout its reactor lifetime and beyond.

Safety analysis. Normal and anticipated abnormal operating conditions must be considered in the analysis of the safety of a reactor or other facility using radioactive material. Hypothetical reactor accidents are analyzed to assess possible consequences and to devise means to prevent or mitigate these consequences.

Operation and test. This function of nuclear engineering is concerned with the supervision and operation of nuclear power reactors and ancillary nuclear facilities.

Affected industries

Nuclear engineers perform these functions for various kinds of employers: (1) architectural engineering firms, in which they handle design, safety analysis, project coordination, construction supervision, quality assurance, quality control, and related matters, (2) reactor vendors and other manufacturing organizations, in which they pursue research, development, design, manufacture, and installation of various components of nuclear systems, (3) electric utility companies, in which they handle planning, construction supervision, reactor-safety analysis, in-core nuclear fuel management, power-reactor economic analysis, environmental-impact assessment, personnel training, plant management, operation-shift supervision, radiation protection, spent-fuel storage, and radioactive-waste management, (4) regulatory agencies, in which they undertake licensing, rule making, safety research, risk analysis, on-site inspection, and research administration, (5) defense programs, in which they are employed in naval and nuclear weapons programs, (6) universities, in which they hold various faculty positions, and (7) national laboratories and industrial research laboratories, in which they carry out advanced research and development on a variety of nuclear programs in nuclear energy areas. Most of the advanced research and development on nuclear-related programs is conducted at national laboratories.

Branches of nuclear engineering. *Nuclear power.* The greatest growth in the nuclear industry has been in the development of nuclear power plants. It is estimated that by the year 2000 one-third of all electric power generated worldwide will come from nuclear power plants.

Nearly all commercial nuclear reactors in operation or under construction are thermal reactors. They are called thermal reactors because their fuel is fissioned by neutrons that have been slowed down by a moderator until they are in thermal equilibrium with the moderator. The boiling water reactor (BWR) and the pressurized water reactor (PWR) are the two predominant types of power reactors in use throughout the world. Both types are called light-water reactors (LWR). The water is used in these reactors as both moderator and coolant. In the BWR, steam is generated by direct boiling of water in the reactor core. In the PWR, steam is produced in an external steam generator rather than in the core, where the coolant under pressure is not allowed to boil. Other types of power reactors include graphite-moderated gas-cooled reactors in use in Great Britain and pressurized heavy-water reactors in Canada.

A major advance in nuclear power is expected with the further development of the liquid-metal fast-breeder reactor (LMFBR). Programs are in progress in several countries to develop and deploy the LMFBR. (The reactor is cooled by a liquid metal, sodium, and fission is caused by fast neutrons. The reactor is called a breeder because it produces more nuclear fuel than it consumes.) Fuel in the breeder is utilized 60 times more effectively than that in light-water reactors. It is estimated that without the breeder the world supply of fissionable material for nuclear power plants could be consumed in a few decades. With the improved fuel utilization provided by the breeder, nuclear power plants would be able to supply the world's electric energy requirement for centuries.

Fusion. Fusion is a potential energy resource with a wide range of applications. The fusion process of combining two light atoms to form a heavier atom, with less mass than the two original atoms, is the basic energy process in the universe (*i.e.,* fusion is the process that takes place in all stars). If fusion can be harnessed for terrestrial applications, the energy can be released in a variety of forms, including charged particles, electromagnetic radiation, and neutrons. Possible applications include electricity production, synthetic fuel production, process-heat applications, and fissile fuel production for fission reactors.

Plasma physics

Fusion research since about 1950 has concentrated on the issues of plasma physics, specifically the production of high-temperature plasmas (100,000,000° C [180,000,000° F] or greater) that can be confined at sufficiently high densities for sufficiently long times to produce net energy. Energy break-even conditions are expected to be demonstrated in several fusion devices in the late 20th century. Fusion physics research has made steady progress, and research efforts have begun to address the important engineering issues of fusion. Among the more important of these issues are those related to extracting useful energy from a plasma and developing complete fuel systems for fusion reactors. These areas are expected to receive increased research and development support in the future.

Naval nuclear propulsion. The use of nuclear reactors to propel naval vessels has revolutionized naval operations throughout the world. The navies of Great Britain, France, the U.S.S.R., China, and the United States are equipped with nuclear-powered ships, which are considered to be of the highest importance to the defense of their countries. Nuclear warships are capable of nearly unlimited high-speed operation without the need of fuel-oil support. In the 25 years following the maiden voyage of the *Nautilus* in

1954, the nuclear navy of the United States steamed more than 80,000,000 kilometres (50,000,000 miles) throughout the oceans of the world, accumulating 25 centuries of reactor-plant operation without any accidents involving a nuclear reactor. By the mid-1980s, more than 40 percent of U.S. combat warships were nuclear-powered.

Nuclear weapons. Fission weapons (atomic bombs), fusion weapons (hydrogen bombs), and combination fission-fusion weapons are part of the world's nuclear arsenal. Nuclear engineers are employed on weapons programs in such diverse activities as research, development, design, fabrication, production, testing, maintenance, and surveillance of a large array of nuclear weapons systems.

Efforts are in progress in the United States to develop, upgrade, and integrate weapons into warhead programs and to explore advanced concepts for future weapons systems. A concept of particular interest is inertial-confinement fusion. This program is directed at determining the feasibility of burning very small pellets of thermonuclear fuel using laser or particle-beam drivers. The program is of interest not only for applications to weapons physics but also for possible energy applications.

Radioisotopes. More than 500 radioisotopes are produced in nuclear reactors. The production, packaging, and application of these isotopes has become a large industry. They are used in heart pacemakers, medical research, sterilization of medical instruments, industrial tracers, X-ray equipment, curing of plastics, preservation of food, and as an energy source in electric generators. Perhaps the most important use of radioisotopes is in the field of medicine. They are used in procedures for half of all patients admitted to hospitals in the United States.

Nuclear-waste management. Nuclear wastes can be classified in two groups, low-level and high-level. Low-level wastes come from nuclear power facilities, hospitals, and research institutions and include such items as contaminated clothing, wiping rags, tools, test tubes, needles, and other medical research materials. In the disposal of low-level wastes, the wastes are reduced in volume, then packaged in leak-proof containers, which are placed in an earth-covered trench in a low-level-waste disposal site. Such sites should be continuously monitored to detect any migration of radioactive material. High-level wastes are highly radioactive and derive from the chemical reprocessing of spent fuel elements and from the weapons program.

By the late 20th century many countries were evaluating potential nuclear-waste disposal sites and developing terminal waste-storage technology. All these countries were preparing to handle high-level wastes. All had identified geologic formations that appeared to be technically feasible for repositories. In 1982 the U.S. Congress passed legislation establishing schedules for the selection, development, licensing, and construction of repositories for the safe, permanent storage of high-level waste. (I.Bo./Ed.)

BIBLIOGRAPHY. Works on the history of the engineering profession include A.P.M. FLEMING and H.J.S. BROCKLEHURST, *A History of Engineering* (1925); R.S. KIRBY *et al., Engineering in History* (1956); and J.K. FINCH, *The Story of Engineering* (1960). Contemporary descriptions may be found in R.J. SMITH, *Engineering as a Career,* 3rd ed. (1969); and T.J. HOOVER and J.C.L. FISH, *The Engineering Profession,* 2nd ed. (1950).

Civil engineering: J.P.M. PANNELL, *An Illustrated History of Civil Engineering* (1964); C.M. NORRIE, *Bridging the Years: A Short History of British Civil Engineering* (1956); H. STRAUB, *A History of Civil Engineering* (Eng. trans. 1953); J.K. FINCH, *Engineering and Western Civilisation* (1951); R.S. KIRBY and P.G. LAURSON, *Early Years of Modern Civil Engineering* (1932); K.L. NASH, *Civil Engineering,* 4th ed. (1967); R. HAMMOND (ed.), *Modern Civil Engineering Practice* (1961); and E.E. MANN, *An Introduction to the Practice of Civil Engineering,* 2nd ed. (1949).

Mechanical engineering: For general reading, see *Engineering Heritage: Highlights from the History of Mechanical Engineering,* 2 vol. (1963–66), issued by the INSTITUTION OF MECHANICAL ENGINEERS. The history of this society is given in R.H. PARSONS, *A History of the Institution of Mechanical Engineers, 1847–1947* (1947).

Chemical engineering: Classic works include G.E. DAVIS, *A Handbook of Chemical Engineering,* 2 vol. (1901, 2nd ed. 1904); and W.H. WALKER, W.K. LEWIS, and W.H. MCADAMS, *The Principles of Chemical Engineering,* 2nd ed. (1927). Current information may be found in *Perry's Chemical Engineers' Handbook,* 6th ed., edited by DON W. GREEN and JAMES O. MALONEY (1984), a comprehensive handbook; H.W. CREMER *et al., Chemical Engineering Practice,* 12 vol. (1956–65), a reference work; and J.M. COULSON and J.F. RICHARDSON, *Chemical Engineering,* 2nd ed., 3 vol. (1964–71), a general textbook.

Electrical and electronics engineering: EDWIN T. LAYTON, *The Revolt of the Engineers* (1971), a discussion of social responsibility in the engineering profession; LIONEL V. BALDWIN and KENNETH S. DOWN, *Educational Technology in Engineering* (1981), a work on study and teaching in engineering; and PHILIP SPORN, *The Social Organization of Electric Power Supply in Modern Societies* (1971), an exposition of the social aspects of engineering for public utilities. ABRAM J. FOSTER, *The Coming of the Electrical Age to the United States* (1979), a history of electrification; JAMES E. BRITTAIN (ed.), *Turning Points in American Electrical History* (1977), a collection of writings on the development of electrical engineering and telecommunications in the United States; BRIAN BOWERS, *A History of Electric Light and Power* (1982), a work that concentrates on electrification in Great Britain; ERNEST BRAUN and STUART MACDONALD, *Revolution in Miniature,* 2nd ed. (1982), a work that explores the impact of semiconductor electronics in industry; and DIRK HANSON, *The New Alchemist* (1982), a work that provides a historical survey of microelectronics in industry.

Petroleum engineering: Standard textbooks are JAMES W. AMYX, D.M. BASS, JR., and R.L. WHITING, *Petroleum Reservoir Engineering* (1960); BENJAMIN C. CRAFT and M.F. HAWKINS, *Applied Petroleum Reservoir Engineering* (1959); EDWARD J. LYNCH, *Formation Evaluation* (1962), on oil-well logging; T.E.W. NIND, *Principles of Oil Well Production* (1964); SYLVAIN J. PIRSON, *Geologic Well Log Analysis* (1983); and LESTER C. UREN, *Petroleum Production Engineering,* 3 vol. (1950–56), on petroleum production economics, oil-field exploitation, and oil-field development. See also ARTHUR W. MCCRAY and FRANK W. COLE, *Oil Well Drilling Technology* (1959); and NATIONAL PETROLEUM COUNCIL, *Impact of New Technology on the U.S. Petroleum Industry: 1946–1965* (1967). *Oil and Gas Journal,* vol. 57, no. 5 (Jan. 28, 1959), is a special issue surveying the first 100 years of the petroleum industry; see also vol. 75, no. 35 (August 1977).

Aerospace engineering: See *Engineering and Technology Enrollments* and *Engineering and Technology Degrees,* two annual publications that supply information on engineering schools and technical education in the United States, both of which are issued by the Engineering Manpower Commission of the American Association of Engineering Societies. FRANK W. ANDERSON, *Orders of Magnitude,* 2nd ed. (1981), a book in the NASA history series, covers the period from 1915 to 1980; JOHN D. ANDERSON, *Introduction to Flight: Its Engineering and History* (1978), deals with theoretical questions of aerodynamics and describes the design and construction of airplanes; and CHARLES H. GIBBS-SMITH, *Flight Through the Ages* (1974), is a survey of aeronautics from its early period to the age of space exploration. TOM D. CROUCH, *A Dream of Wings* (1981), traces the history of aeronautics in the United States; JEROME LEDERER, "Highlights in the Development of Civilian Aircraft," *Automotive Engineering* 88(12):33–43 (December 1980), is a review that dwells on the prominent technical concepts and development of civil air transportation; BARNES W. MCCORMICK, *Aerodynamics, Aeronautics, and Flight Mechanics* (1979), is an illustrated monograph; LELAND M. NICOLAI, *Fundamentals of Aircraft Design* (1975), is an illustrated source, with bibliographies; and RICHARD S. SHEVELL, *Fundamentals of Flight* (1983), is a monograph, with bibliography.

Nuclear engineering: HENRY DEWOLF SMYTH, *Atomic Energy for Military Purposes: The Official Report on the Development of the Atomic Bomb Under the Auspices of the United States Government, 1940–1945* (1945, reprinted 1978), commonly known as the Smyth report, issued by the Manhattan District of the U.S. Corps of Engineers; SAMUEL GLASSTONE and MILTON C. EDLUND, *The Elements of Nuclear Reactor Theory* (1952), a discussion of general principles of reactor technology; HAROLD ETHERINGTON, *Nuclear Engineering Handbook* (1958), a standard reference work that covers all phases of nuclear engineering; SAMUEL GLASSTONE and RALPH H. LOVBERG, *Controlled Thermonuclear Reactions* (1960, reprinted 1975), an introduction to nuclear fusion; SAMUEL GLASSTONE and ALEXANDER SESONSKE, *Nuclear Reactor Engineering,* 3rd ed. (1981), a basic work for nuclear engineering education; GLENN THEODORE SEABORG and WILLIAM R. CORLISS, *Man and Atom* (1971), a work that explores peaceful applications of nuclear energy; RICHARD G. HEWLETT and FRANCIS DUNCAN, *Nuclear Navy, 1946–1962* (1974), a historical and analytical study; RAYMOND LEROY MURRAY, *Nuclear Energy,* 2nd ed. (1980), an introduction to basic nuclear processes; and MANSON BENEDICT, THOMAS H. PIGFORD, and HANS WOLFGANG LEVI, *Nuclear Chemical Engineering,* 2nd ed. (1981), a comprehensive work that emphasizes nuclear processing and includes useful information on metallurgy.

English Literature

Although for the purposes of this article English literature is treated as being confined to writings in English by natives or inhabitants of the British Isles (including Ireland), it is to a certain extent the case that literature—and this is particularly true of the literature written in English—knows no frontiers. Thus, English literature can be regarded as a cultural whole of which the mainstream literatures of the United States, Australia, New Zealand, and Canada and important elements in the literatures of other Commonwealth or ex-Commonwealth countries are parts (see AMERICAN LITERATURE; AUSTRALIA AND NEW ZEALAND, LITERATURES OF; and CANADIAN LITERATURE).

English literature has sometimes been stigmatized as insular. It can be argued that no single English novel attains the universality of the Russian writer Leo Tolstoy's *War and Peace* or the French writer Gustave Flaubert's *Madame Bovary*. Yet in the Middle Ages the Old English literature of the subjugated Saxons was leavened by the Latin and Anglo-Norman French writings, eminently foreign in origin, in which the churchmen and the Norman conquerors expressed themselves. From this combination emerged a flexible and subtle linguistic instrument exploited by Geoffrey Chaucer and brought to supreme application by William Shakespeare. During the Renaissance the renewed interest in classical learning and values had an important effect on English literature, as on all of the arts; and ideas of Augustan literary propriety in the 18th century and reverence in the 19th century for a less specific, though still selectively viewed, classical antiquity continued to shape the literature. All three of these impulses derived from a foreign source, namely the Mediterranean basin. The Decadents of the late 19th century and modernists of the early 20th looked to continental European individuals and movements for inspiration. Nor was attraction toward European intellectualism dead in the late 20th century, for by the mid-1980s the approach known as structuralism, a phenomenon predominantly French and German in origin, infused the very study of English literature itself in a host of published critical studies and university departments.

Further, Britain's past imperial glories around the globe, particularly those that were connected with the Indian subcontinent, continued to inspire literature—in some cases wistful, in other cases hostile. Finally, English literature has enjoyed a certain diffusion abroad, not only in predominantly English-speaking countries but also in all those others where English is the first choice of study as a second language.

English literature is therefore not so much insular as detached from the continental European tradition across the Channel. It is strong in all the conventional categories of the bookseller's list: in Shakespeare it has a dramatist of world renown; in poetry, a genre notoriously resistant to adequate translation and therefore difficult to compare with the poetry of other literatures, it is so peculiarly rich as to merit inclusion in the front rank; English literature's humour has been found as hard to convey to foreigners as poetry, if not more so—a fact at any rate permitting bestowal of the label "idiosyncratic"; English literature's remarkable body of travel writings constitutes another counterthrust to the charge of insularity; in autobiography, biography, and historical writing English literature compares with the best of any culture; and children's literature, fantasy, essays, and journals, which tend to be considered minor genres, are all fields of exceptional achievement as regards English literature. Even in philosophical writings, popularly thought of as hard to combine with literary value, thinkers such as Thomas Hobbes, John Locke, David Hume, John Stuart Mill, and Bertrand Russell stand comparison for lucidity and grace with the best of the French philosophers and the masters of classical antiquity.

Some of English literature's most distinguished practitioners in the 20th century—from Henry James and Joseph Conrad at its beginning to V.S. Naipaul and Tom Stoppard more recently—were of foreign origin. What is more, none of the aforementioned had as much in common with his adoptive country as did, for instance, Doris Lessing and Peter Porter (two other distinguished writer-immigrants to Britain) by virtue both of having been born into a British family and of having been brought up on British Commonwealth soil.

On the other hand, during the same period in the 20th century, many notable practitioners of English literature left Britain to live abroad: James Joyce, D.H. Lawrence, Aldous Huxley, Christopher Isherwood, Robert Graves, Graham Greene, Muriel Spark, Anthony Burgess, and Sir Angus Wilson. In one case, that of Samuel Beckett, this process was carried to the extent of writing works first in French and then translating them into English.

Even English literature considered purely as a product of the British Isles is extraordinarily heterogeneous, however. Literature actually written in those Celtic tongues once prevalent in Cornwall, Ireland, Scotland, and Wales—called the "Celtic Fringe"—is treated separately (see CELTIC LITERATURE). Yet Irish, Scots, and Welsh writers have contributed enormously to English literature even when they have written in dialect, as the 18th-century poet Robert Burns and the 20th-century Scots writer Alasdair Gray have done. In the latter half of the 20th century interest began also to focus on writings in English or English dialect by recent settlers in Britain, such as Afro-Caribbeans and people from Africa proper, the Indian subcontinent, and East Asia.

Even within England, culturally and historically the dominant partner in the union of territories comprising Britain, literature has been as enriched by strongly provincial writers as by metropolitan ones. Another contrast more fruitful than not for English letters has been that between social milieus, however much observers of Britain in their own writings may have deplored the survival of class distinctions. As far back as medieval times a courtly tradition in literature cross-fertilized with an earthier demotic one. Shakespeare's frequent juxtaposition of royalty in one scene with plebeians in the next reflects a very British way of looking at society. This awareness of differences between high life and low, a state of affairs fertile in creative tensions, is observable throughout the history of English literature.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, Part Six, Division II, Section 621. (Ed.)

This article is divided into the following sections:

The Old English period

POETRY

The Angles, Saxons, and Jutes who invaded Britain in the 5th and 6th centuries brought with them the common Germanic metre; but of their earliest oral poetry, probably used for panegyric, magic, and short narrative, little or none survives. For nearly a century after the conversion of King Aethelberht I of Kent to Christianity in 597, there is no evidence that the English wrote poetry in their own language. But St. Bede the Venerable, in his *Historia ecclesiastica gentis Anglorum* ("Ecclesiastical History of the English People"), wrote that in the late 7th century Caedmon, an illiterate Northumbrian cowherd, was inspired in a dream to compose a short hymn in praise of the creation. Caedmon later composed verses based on Scripture, which was expounded for him by monks at Streaneshalch (Whitby), but only the "Hymn of Creation" survives. Caedmon legitimized the native verse form by adapting it to Christian themes. Others, following his example, gave England a body of vernacular poetry unparalleled in Europe before the end of the 1st millennium.

Bede's account of Caedmon

Alliterative verse. Virtually all Old English poetry is written in a single metre, a four-stress line with a syntactical break, or caesura, between the second and third stresses, and with alliteration linking the two halves of the line; this pattern is occasionally varied by six-stress lines. The poetry is formulaic, drawing on a common set of stock phrases and phrase patterns, applying standard epithets to various classes of characters, and depicting scenery with such recurring images as the eagle and wolf, which wait during battles to feast on carrion, and the ice and snow, which appear in the landscape to signal sorrow. In the best poems such formulas, far from being tedious, give a strong impression of the richness of the cultural fund from which poets could draw. Other standard devices of this poetry are the kenning, a metaphorical name for a thing, usually expressed in a compound noun (*e.g.,* "swan-road" used to name the sea); and variation, the repeating of a single idea in different words, with each repetition adding a new level of meaning. That these verse techniques changed little during 400 years of literary production suggests the extreme conservatism of Anglo-Saxon culture.

The major manuscripts. Most Old English poetry is preserved in four manuscripts of the late 10th and early 11th centuries. The Beowulf manuscript (British Library) contains *Beowulf, Judith,* and three prose tracts; the Exeter Book (Exeter cathedral) is a miscellaneous gathering of lyrics, riddles, didactic poems, and religious narratives; the Junius manuscript (Bodleian Library, Oxford) contains biblical paraphrases; and the Vercelli Book (cathedral library, Vercelli, Italy) contains saints' lives, several short religious poems, and prose homilies. In addition to the poems in these books are historical poems in the *Anglo-Saxon Chronicle;* poetic renderings of Psalms 51–150; the 31 "Metres" included in King Alfred the Great's translation of Boethius' *Consolation of Philosophy;* magical, didactic, elegiac, and heroic poems; and others, miscellaneously interspersed with prose, jotted in margins and even worked in stone or metal.

Problems of dating. Few poems can be dated as closely as Caedmon's "Hymn." King Alfred's compositions fall into the late 9th century, and Bede composed his "Death Song" within 50 days of his death on May 25, 735. His-

torical poems like "The Battle of Brunanburh" (after 937) and "The Battle of Maldon" (after 991) are fixed by the dates of the events they commemorate. A translation of one of Aldhelm's riddles is found not only in the Exeter Book but also in an early 9th-century manuscript at Leiden; it can be no later than the Leiden manuscript. And at least a part of "The Dream of the Rood" can be dated by an excerpt carved on the 8th-century Ruthwell Cross (in Dumfriesshire, Scotland). But in the absence of such indications, Old English poems are hard to date, and the scholarly consensus that most were composed in the Midlands and the North in the 8th and 9th centuries has crumbled in recent years. Many now hold that "The Wanderer," *Beowulf,* and other poems once assumed to be early are of the 9th or 10th century. For most poems, little more than that they were written between the 8th and the 11th centuries can be said with certainty.

Religious verse. If few poems can be dated accurately, still fewer can be attributed to particular poets. The most important author from whom a considerable body of work survives is Cynewulf, who wove his runic signature into the epilogues of four poems. Aside from his name, little is known of him; he probably lived in the 9th century in Mercia. His works include *The Fates of the Apostles,* a short martyrology; *The Ascension* (also called *Christ II*), a homily and biblical narrative; *Juliana,* a saint's passion set in the reign of Maximian (late 3rd century AD); and *Elene,* perhaps the best of his poems, which describes the mission of St. Helena, mother of the emperor Constantine, to recover Christ's cross. Cynewulf's work is lucid and technically elegant; his theme is the continuing evangelical mission from the time of Christ to the triumph of Christianity under Constantine. Several poems not by Cynewulf are associated with him because of their subject matter. These include two lives of St. Guthlac and *Andreas,* the story of St. Andrew among the Mermedonians, which has stylistic affinities with *Beowulf.* Also in the "Cynewulf group" are several poems with Christ as their subject, of which the most important is "The Dream of the Rood," in which the cross speaks of itself as Christ's loyal thane and yet the instrument of his death. This tragic paradox echoes a recurring theme of secular poetry and at the same time movingly expresses the religious paradoxes of Christ's triumph in death and mankind's redemption from sin.

Cynewulf's poems

The Old Testament narratives (*Genesis, Exodus,* and *Daniel*) of the Junius manuscript were once attributed to Caedmon but now are thought to be of anonymous authorship. Of these *Exodus* is remarkable for its intricate diction and bold imagery. The fragmentary *Judith* of the Beowulf manuscript stirringly embellishes the story from the apocrypha of the heroine who led the Jews to victory over the Assyrians.

Elegiac and heroic verse. The term elegy is used of Old English poems that lament the loss of worldly goods, glory, or human companionship. "The Wanderer" is narrated by a man, deprived of lord and kinsmen, whose journeys lead him to the realization that there is stability only in heaven. "The Seafarer" is similar, but its journey motif more explicitly symbolizes the speaker's spiritual yearnings. Several others have similar themes, and three elegies, "The Husband's Message," "The Wife's Lament," and "Wulf and Eadwacer," describe what appears to be a conventional situation: the separation of husband and wife by the husband's exile.

"Deor" bridges the gap between the elegy and the heroic poem, for in it a poet laments the loss of his position at court by alluding to sorrowful stories from Germanic legend. *Beowulf* itself narrates the battles of Beowulf, a prince of the Geats (a tribe in what is now southern Sweden), against the monstrous Grendel, Grendel's mother, and a fire-breathing dragon. The account contains some of the best elegiac verse in the language; and by setting marvelous tales against a historical background in which victory is always temporary and strife is always renewed, the poet gives the whole an elegiac cast. *Beowulf* also is one of the best religious poems, not only because of its explicitly Christian passages but also because Beowulf's monstrous foes are depicted as God's enemies and Beowulf himself as God's champion. Other heroic narratives are fragmentary.

Beowulf

Of "The Battle of Finnsburh" and "Waldere" only enough remains to indicate that when whole they must have been fast paced and stirring.

Of several poems dealing with English history and preserved in the *Anglo-Saxon Chronicle,* the most notable is "The Battle of Brunanburh," a panegyric on the occasion of King Athelstan's victory over a coalition of Norsemen and Scots in the year 937. But the best historical poem is not from the *Chronicle.* "The Battle of Maldon," which describes the defeat of Aldorman Byrhtnoth at the hands of Viking invaders in 991, states eloquently the heroic ideal, contrasting the determination of some of Byrhtnoth's thanes to avenge his death or die in the attempt with the cowardice of others who left the field. Minor poetic genres include catalogs (two sets of "Maxims" and "Widsith," a list of rulers, tribes, and notables in the heroic age), dialogues, metrical prefaces and epilogues to prose works of the Alfredian period, and liturgical poems associated with the Benedictine Office.

PROSE

The earliest English prose work, the law code of King Aethelberht I of Kent, was written within a few years of St. Augustine of Canterbury's arrival in England (597). Other 7th- and 8th-century prose, similarly practical in character, includes more laws, wills, and charters. According to Cuthbert, who was a monk at Jarrow, Bede had just finished a translation of the Gospel of St. John at the time of his death, though this does not survive; and two medical tracts, a *Herbarium* and *Medicina de quadrupedibus,* very likely date from the 8th century.

Early translations into English. But the earliest literary prose dates from the late 9th century, when King Alfred, eager to improve the state of English learning, led a vigorous program to translate into English "certain books that are necessary for all men to know." Alfred himself translated St. Gregory I the Great's *Pastoral Care,* Boethius' *Consolation of Philosophy,* St. Augustine of Hippo's *Soliloquies* and the first 50 psalms. His *Pastoral Care* is a fairly literal translation, but his Boethius is extensively restructured and revised to make explicit the Christian message that medieval commentators saw in that work. He revised the *Soliloquies* even more radically, departing from his source to draw from St. Jerome, Gregory, and other works by Augustine. Alfred's prefaces to these works are of great historical interest.

At Alfred's urging Bishop Werferth of Worcester translated the *Dialogues* of Gregory; probably Alfred also inspired anonymous scholars to translate Bede's *Historia ecclesiastica* and Paulus Orosius' *Historiarum adversum paganos libri vii* ("History Opposing the Pagans, In Seven Books"). Both of these works are much abridged; the Bede translation follows its source slavishly, but the translator of Orosius added many details of northern European geography and also accounts of the voyages of Ohthere the Norwegian and Wulfstan the Dane. These accounts, in addition to their geographical interest, show that friendly commerce between England and Scandinavia was possible even during the Danish wars. The *Anglo-Saxon Chronicle* probably originated in Alfred's reign. Its earliest annals (from 60 BC) are laconic, except the entry for 755, which records in detail a feud between the West Saxon king Cynewulf and the would-be usurper Cyneheard. The entries covering the Danish wars of the late 9th century are much fuller, and those running from the reign of Aethelred I to the Norman Conquest in 1066 (when the *Chronicle* exists in several versions) contain many passages of excellent writing. The early 10th century is not notable for literary production, but some of the homilies in the Vercelli Book and the Blickling manuscript (Scheide Library, Princeton University) may belong to that period.

The *Anglo-Saxon Chronicle*

Late 10th- and 11th-century prose. The Benedictine reform of the mid-10th century brought about a period of lively literary activity. Aethelwold, bishop of Winchester and one of the leaders of the reform, translated the Rule of St. Benedict. But the greatest and most prolific writer of this period was his pupil Aelfric, abbot of Eynsham, whose works include three cycles of 40 homilies each (*Catholic Homilies,* 2 vol., and the *Lives of the Saints*), as well as

homilies not in these cycles; a Latin grammar; a treatise on science; pastoral letters; and several translations. His Latin *Colloquy,* supplied with an Old English version by an anonymous glossarist, gives a charming picture of everyday life in Anglo-Saxon England. Aelfric wrote with lucidity and astonishing beauty, using the rhetorical devices of Latin literature frequently but without ostentation; his later alliterative prose, which loosely imitates the rhythms of Old English poetry, influenced writers long after the Norman Conquest. Wulfstan, archbishop of York, wrote legal codes, both civil and ecclesiastical, and a number of homilies, including *Sermo Lupi ad Anglos* ("Wulf's Address to the English"), a ferocious denunciation of the morals of his time. To judge from the number of extant manuscripts, these two writers were enormously popular. Byrhtferth of Ramsey wrote several Latin works and the *Enchiridion,* a textbook on the calendar, notable for its ornate style. Numerous anonymous works, some of very high quality, were produced in this period, including homilies, saints' lives, dialogues, and translations of such works as the Gospels, several Old Testament books, liturgical texts, monastic rules, penitential handbooks, and the romance *Apollonius of Tyre* (translated from Latin but probably derived from a Greek original). The works of the monastic reform were written during a few remarkable decades around the turn of the millennium. Little original work can be securely dated to the period after Wulfstan's death (1023), but the continued vigour of the *Anglo-Saxon Chronicle* shows that good Old English prose was written right up to the Norman Conquest. By the end of this period English had been established as a literary language with a polish and versatility unequaled among European vernaculars.

The Early Middle English period

POETRY

The Norman Conquest worked no immediate transformation on either the language or literature of the English. Older poetry continued to be copied during the last half of the 11th century; two poems of the early 12th century—"Durham," which praises that city's cathedral and its relics, and "Instructions for Christians," a didactic piece—show that correct alliterative verse could be composed well after 1066. But even before the Conquest rhyme had begun to supplant rather than supplement alliteration in some poems, which continued to use the older four-stress line but the rhythms of which varied from the set types used in classical Old English verse. A post-Conquest example is "The Grave," which contains several rhyming lines; a poem from the *Anglo-Saxon Chronicle* on the death of William the Conqueror, lamenting his cruelty and greed, has more rhyme than alliteration.

Influence of French poetry. By the end of the 12th century English poetry had been so heavily influenced by French models that such a work as the long epic *Brut* (*c.* 1200) by Lawamon, a Worcestershire priest, seems archaic for mixing alliterative lines with rhyming couplets while generally eschewing French vocabulary. The *Brut* mainly draws upon Wace's Anglo-Norman *Roman de Brut* (1155; based in turn upon Geoffrey of Monmouth's *Historia regum Britanniae,* or *History of the Kings of Britain*), but in Lawamon's hands the Arthurian story takes on a Germanic and heroic flavour largely missing in Wace. The *Brut* exists in two manuscripts, one written shortly after 1200 and the other some 50 years later. That the later version has been extensively modernized and somewhat abridged suggests the speed with which English language and literary tastes were changing in this period. *The Proverbs of Alfred* also were written in the late 12th century; these deliver conventional wisdom in a mixture of rhymed couplets and alliterative lines, and it is hardly likely that any of the material they contain actually originated with the king whose wisdom they celebrate. The early 13th-century *Bestiary* mixes alliterative lines, three- and four-stress couplets, and septenary lines, but the logic behind this mix is more obvious than in the *Brut* and the *Proverbs,* for the poet was imitating the varied metres of his Latin source. More regular in form than these poems is the anonymous *Poema morale* in septenary couplets, in which an old man delivers a dose of moral advice to his presumably younger audience.

Lawamon's *Brut*

By far the most brilliant poem of this period is *The Owl and the Nightingale* (written after 1189), an example of the popular debate genre. The two birds argue topics ranging from their hygienic habits, looks, and songs to marriage, prognostication, and the proper modes of worship. The nightingale stands for the joyous aspects of life, the owl for the sombre; there is no clear winner, but the debate ends as the birds go off to state their cases to one Nicholas of Guildford, a wise man. The poem is learned in the clerical tradition but wears its learning lightly as the disputants speak in colloquial and sometimes earthy language. Like the *Poema morale, The Owl and the Nightingale* is metrically regular (octosyllabic couplets), but it uses the French metre with an assurance that is astonishing in so early a poem.

Didactic poetry. The 13th century saw a rise in the popularity of long didactic poems presenting biblical narrative, saints' lives, or moral instruction for those untutored in Latin or French. The most idiosyncratic of these is the *Ormulum* by Orm, an Augustinian canon in the north of England. Written in some 20,000 lines arranged in unrhymed but metrically rigid couplets, the work is interesting mainly in that the manuscript that preserves it is Orm's autograph and shows his somewhat fussy (and ineffectual) efforts to reform and regularize English spelling. Other biblical paraphrases are *Genesis and Exodus, Jacob and Joseph,* and the vast *Cursor mundi,* whose subject, as its title suggests, is the whole history of the world. An especially popular work was the *South English Legendary,* which began as a miscellaneous collection of saints' lives but was expanded by later redactors and rearranged in the order of the church calendar. The didactic tradition continued into the 14th century with Robert Mannyng's *Handlyng Synne,* a confessional manual the expected dryness of which is relieved by the insertion of lively narratives, and the *Pricke of Conscience,* a summary of theology sometimes attributed to Richard Rolle.

Verse romance. The earliest examples of verse romance, a genre that would remain popular through the Middle Ages, appeared in the 13th century. *King Horn* and *Floris and Blauncheflour* both are preserved in a manuscript of around 1250. *King Horn,* oddly written in short two- and three-stress lines, is a vigorous tale of a kingdom lost and regained, with a subplot concerning Horn's love for Princess Rymenhild. *Floris and Blauncheflour* is more exotic, being the tale of a pair of royal lovers who become separated and, after various adventures in eastern lands, reunited. Not much later than these is *The Lay of Havelok the Dane,* a tale of princely love and adventure similar to *King Horn* but more competently executed. Many more such romances were produced in the 14th century. Popular subgenres were "the matter of Britain" (Arthurian romances such as *Of Arthour and of Merlin* and *Ywain and Gawain*); "the matter of Troy" (tales of antiquity such as *The Seege of Troye* and *Kyng Alisaunder*); and the English Breton lays, stories of otherworldly magic, such as *Lai le Freine* and *Sir Orfeo,* modeled after those of professional Breton storytellers. These relatively unsophisticated works were no doubt written for a bourgeois audience, and the manuscripts that preserve them are early examples of commercial book production. The humorous beast epic makes its first appearance in the 13th century in *The Fox and the Wolf,* taken indirectly from the Old French *Roman de Renart.* In the same manuscript with this work is *Dame Sirith,* the earliest English fabliau. Another sort of humour is found in *The Land of Cockaygne,* which depicts a utopia better than heaven, where rivers run with oil, milk, honey, and wine, geese fly about already roasted, and monks hunt with hawks and dance with nuns.

Floris and Blauncheflour

The lyric. The lyric was virtually unknown to Old English poets: poems like "Deor" and "Wulf and Eadwacer," which have been called lyrics, are thematically different from those that began to circulate orally in the 12th century and to be written down in great numbers in the 13th; and these Old English poems have a stronger narrative component than the later productions. The most frequent

topics in the Middle English secular lyric are springtime and romantic love; many rework such themes tediously, but some, such as "Foweles in the frith" (13th century) and "Ich am of Irlaunde" (14th century), convey strong emotions in a few lines. Two lyrics of the early 13th century, "Mirie it is while sumer ilast" and "Sumer is icumen in," are preserved with musical settings, and probably most of the others were meant to be sung. The dominant mood of the religious lyrics is passionate: the poets sorrow for Christ on the Cross and for Mary, celebrate the "five joys" of Mary, and import language from love poetry to express religious devotion. Excellent early examples are "Nou goth sonne under wod" and "Stond wel, moder, ounder rode." Many of the lyrics are preserved in manuscript anthologies, of which the best is British Library manuscript Harley 2253 from the early 14th century. The love poems in this collection, such as "Alysoun" and "Blow, Northerne Wynd," take after the poems of the Provençal troubadours but are less formal and abstract and therefore more lively. The religious lyrics also are of high quality; but the most remarkable of the Harley Lyrics, "The Man in the Moon," far from being about love or religion, imagines the man in the Moon as a simple peasant, sympathizes with his hard life, and offers him some useful advice on how to best the village hayward.

A poem such as "The Man in the Moon" serves as a reminder that, although the poetry of the early Middle English period is increasingly influenced by the Anglo-Norman literature produced for the courts, it is seldom "courtly." Most English poets, whether writing about kings or peasants, looked at life from a middle-class perspective. If their work sometimes lacks sophistication, it nevertheless has a vitality that comes from preoccupation with daily affairs; its practicality, as much as its language, gives it a distinctly English flavour.

PROSE

Old English prose texts were copied for more than a century after the Norman Conquest; the homilies of Aelfric were especially popular, and King Alfred's translations of Boethius and Augustine survive only in 12th-century manuscripts. In the early 13th century an anonymous worker at Worcester supplied glosses to certain words in a number of Old English manuscripts, demonstrating that by this time the older language was beginning to pose difficulties for readers.

The composition of English prose also continued without interruption. Two manuscripts of the *Anglo-Saxon Chronicle* exhibit very strong prose for years after the Conquest, and one of these, *The Peterborough Chronicle,* continues to the year 1154. Two manuscripts of around 1200 contain 12th-century sermons, and another has a workmanlike compilation on the "Vices and Virtues," composed around 1200. But the English language faced stiff competition from both Anglo-Norman (the insular dialect of French being used increasingly in the monasteries) and Latin, a language intelligible to speakers of both English and French. It was inevitable, then, that the production of English prose should decline in quantity, if not in quality. The great prose works of this period were composed mainly for those who could read only English—women especially. In the West Midlands the Old English alliterative prose tradition remained very much alive into the 13th century, when the several texts known collectively as the Katherine Group were written. "St. Katherine," "St. Margaret," and "St. Juliana," found together in a single manuscript, have rhythms strongly reminiscent of those of Aelfric and Wulfstan. So, to a lesser extent, do "Hali Meithhad" ("Holy Maidenhood") and "Sawles Warde" ("The Guardianship of the Soul") from the same book, but newer influences can be seen in these works as well: as the title of another devotional piece, "The Wohunge of Ure Lauerd" ("The Wooing of Our Lord"), suggests, the prose of this time often has a rapturous, even sensual flavour, and, like the poetry, it frequently employs the language of love to express religious fervour.

Composition in English, Anglo-Norman, and Latin

Further removed from the Old English prose tradition, though often associated with the Katherine Group, is the *Ancrene Wisse* ("Guide for Anchoresses," also known as the *Ancrene Riwle,* or "Rule for Anchoresses"), a manual for the guidance of women recluses outside the regular orders. This anonymous work, which was translated into French and Latin and remained popular until the 16th century, is notable for its humanity, practicality, and insight into human nature but even more for its brilliant style. Like the other prose of its time, it uses alliteration as ornament, but it is more indebted to new fashions in preaching, which had originated in the universities, than to native traditions. With its richly figurative language, rhetorically crafted sentences, and carefully logical divisions and subdivisions, it manages to achieve in English the effects that such contemporary writers as John of Salisbury and Walter Map were striving for in Latin.

Little noteworthy prose was written in the late 13th century. In the early 14th century Dan Michel produced in Kentish the *Ayenbite of Inwit* ("Prick of Conscience"), a translation from French. But the best prose of this time is by the mystic Richard Rolle, the hermit of Hampole, whose English tracts include *The Commandment, Meditations on the Passion,* and *The Form of Perfect Living,* among others. His intense and stylized prose was among the most popular of the 14th century and inspired such later works as Walter Hilton's *Scale of Perfection,* Julian of Norwich's *Sixteen Revelations of Divine Love,* and the anonymous *Cloud of Unknowing.* (P.S.Ba.)

The later Middle English and early Renaissance periods

One of the most important factors in the nature and development of English literature between about 1350 and 1550 was the peculiar linguistic situation in England at the beginning of the period. Among the small minority of the population that could be regarded as literate, bilingualism and even trilingualism were common. Insofar as it was considered a serious literary medium at all, English was obliged to compete on uneven terms with Latin and with the Anglo-Norman dialect of French widely used in England at the time. Moreover, extreme dialectal diversity within English itself made it difficult for vernacular writings, irrespective of their literary pretensions, to circulate very far outside their immediate areas of composition, a disadvantage not suffered by writings in Anglo-Norman and Latin. Literary culture managed to survive and in fact to flourish in the face of such potentially crushing factors as the catastrophic mortality of the Black Death (1347–51), chronic external and internal military conflicts in the form of the Hundred Years' War and the Wars of the Roses, and serious social, political, and religious unrest, as evinced in the Peasants' Revolt (1381) and the rise of Lollardism (centred on the religious teachings of John Wycliffe). All the more remarkable then was the literary and linguistic revolution that took place in England between about 1350 and 1400 and that was slowly and soberly consolidated over the subsequent 150 years.

LATER MIDDLE ENGLISH POETRY

The revival of alliterative poetry. The most puzzling episode in the development of later Middle English literature was the apparently sudden reappearance of unrhymed alliterative poetry in the mid-14th century. Debate continues as to whether the group of long, serious, and sometimes learned poems written between about 1350 and the first decade of the 15th century should be regarded as an "alliterative revival" or rather as the late flowering of a largely lost native tradition stretching back to the Old English period. The earliest examples of the phenomenon, *William of Palerne* and *Winner and Waster,* are both datable to the 1350s, but neither poem exhibits to the full all the characteristics of the slightly later poems central to the movement. *William of Palerne,* condescendingly commissioned by a nobleman for the benefit of "them that know no French," is a homely paraphrase of a courtly continental romance, the only poem in the group to take love as its central theme. The poet's technical competence in handling the difficult syntax and diction of the alliterative style is not, however, to be compared with that of *Winner and Waster*'s author, who exhibits full mastery of the

form, particularly in brilliant descriptions of setting and spectacle. This poem's topical concern with social satire links it primarily with another, less formal body of alliterative verse, of which William Langland's *Piers Plowman* was the principal representative and exemplar. Indeed, *Winner and Waster,* with its sense of social commitment and occasional apocalyptic gesture, may well have served as a source of inspiration for Langland himself.

The expression alliterative revival should not be taken to imply a return to the principles of classical Old English versification. The authors of the later 14th-century alliterative poems either inherited or developed their own conventions, which resemble those of the Old English tradition in only the most general way. The syntax and particularly the diction of later Middle English alliterative verse were also distinctive, and the search for alliterating phrases and constructions led to the extensive use of archaic, technical, and dialectal words. Hunts, feasts, battles, storms, and landscapes were described with a brilliant concretion of detail rarely paralleled since, while the abler poets also contrived subtle modulations of the staple verse-paragraph to accommodate dialogue, discourse, and argument. Among the poems central to the movement were three pieces dealing with the life and legends of Alexander, the massive *Destruction of Troy,* and the *Siege of Jerusalem.* The fact that all of these derived from various Latin sources suggests that the anonymous poets were likely to have been clerics with a strong, if bookish, historical sense of their romance "matters." The "matter of Britain" was represented by an outstanding composition, the alliterative *Morte Arthure,* an epic portrayal of King Arthur's conquests in Europe and his eventual fall, combining a strong narrative thrust with considerable density and subtlety of diction. A gathering sense of inevitable transitoriness gradually tempers the virile realization of heroic idealism, and it is not surprising to find that the poem was later used by Sir Thomas Malory as a source for his prose account, *Le Morte Darthur* (completed *c.* 1470).

The alliterative movement would today be regarded as a curious but inconsiderable episode, were it not for four other poems now generally attributed to a single anonymous author: the chivalric romance *Sir Gawayne and the Grene Knight,* two homiletic poems called *Patience* and *Purity* (or *Cleanness*), and an elegiac dream vision known as *Pearl,* all miraculously preserved in a single manuscript dated *c.* 1400. The poet of *Sir Gawayne* far exceeded the other alliterative writers in his mastery of form and style, and though he wrote ultimately as a moralist, human warmth and sympathy (often taking comic form) were also close to the heart of his work. *Patience* relates the biblical story of Jonah as a human comedy of petulance and irascibility set off against God's benign forbearance. *Purity* imaginatively re-creates several monitory narratives of man's impurity and its consequences in a spectacular display of poetic skill: the Flood, the destruction of Sodom, and Belshazzar's Feast. The poet's principal achievement, however, was *Sir Gawayne,* in which he used the conventional apparatus of chivalric romance to engage in a serious exploration of man's moral conduct in the face of the unknown. The hero, a questing knight of Arthur's court, embodies a combination of the noblest chivalric and spiritual aspirations of the age, but instead of triumphing in the conventional way, he fails when tested (albeit rather unfairly) by mysterious supernatural powers. No paraphrase can hope to recapture the brilliant imaginative resources displayed in the telling of the story and the structuring of the poem as a work of art. The *Pearl* stands somewhat aside from the alliterative movement proper. In common with a number of other poems of the period, it was composed in stanzaic form, with alliteration used for ornamental effect. Technically it is one of the most complex poems in the language, an attempt to work in words an analogy to the jeweler's art. The jeweler-poet is vouchsafed a heavenly vision in which he sees his pearl, the discreet symbol used in the poem for a lost infant daughter who has died to become a bride of Christ. She offers theological consolation for his grief, expounding the way of salvation and the place of human life in a transcendental and extra-temporal view of things.

Sir Gawayne and the Grene Knight

The alliterative movement was primarily confined to poets writing in northern and northwestern England, who showed little regard for courtly, London-based literary developments. It is likely that alliterative poetry, under aristocratic patronage, filled a gap in the literary life of the provinces caused by the decline of Anglo-Norman in the latter half of the 14th century. Alliterative poetry was not unknown in London and the southeast, but it penetrated those areas in a modified form and in poems that dealt with different subject matter.

Piers Plowman

William Langland's long alliterative poem *Piers Plowman* begins with a vision of the world seen from the Malvern Hills in Worcestershire, where, tradition has it, the poet was born and brought up, and where he would have been open to the influence of the alliterative movement. If what he tells about himself in the poem is true (and there is no other source of information), he later lived obscurely in London as an unbeneficed cleric. Langland wrote in the unrhymed alliterative mode, but he modified it in such a way as to make it more accessible to a wider audience by treating the metre more loosely and avoiding the arcane diction of the provincial poets. His poem exists in three versions: A, *Piers Plowman* in its short, early form, dating from the 1360s; B, a major revision and extension of A made in the late 1370s; and C (1380s), a less "literary" version of B, apparently intended to bring its doctrinal issues into clearer focus. The poem takes the form of a series of dream visions dealing with the social and spiritual predicament of later 14th-century England against a sombre apocalyptic backdrop. Realistic and allegorical elements are mingled in a phantasmagoric way, and both the poetic medium and the structure are frequently subverted by the writer's spiritual and didactic impulses. Passages of involuted theological reasoning mingle with scatological satire, and moments of sublime religious feeling appear alongside forthright political comment. This makes it a work of the utmost difficulty, defiant of categorization, but at the same time Langland never fails to convince the reader of the passionate integrity of his writing. His bitter attacks on political and ecclesiastical corruption (especially among the friars) quickly struck chords with his contemporaries. Among minor poems in the same vein were *Mum and the Sothsegger* (*c.* 1399–1406) and a Lollard piece called *Pierce the Ploughman's Creed* (*c.* 1395). In the 16th century *Piers Plowman* was issued as a printed book and was used for apologetic purposes by the early Protestants.

Courtly poetry. Apart from a few late and minor reappearances in Scotland and the northwest of England, the alliterative movement was over before the first quarter of the 15th century had passed. The other major strand in the development of English poetry from about 1350 proved much more durable. The cultivation and refinement of human sentiment with respect to love, already present in earlier 14th-century writings such as the Harley Lyrics, took firm root in English court culture during the reign of Richard II (1377–99). English began to displace Anglo-Norman French as the language spoken at court and in aristocratic circles, and signs of royal and noble patronage for English vernacular writers became evident. These processes undoubtedly created some of the conditions in which a writer of Chaucer's interests and temperament might flourish, but they were encouraged and given direction by his genius in establishing English as a literary language.

Chaucer and Gower. Geoffrey Chaucer, a Londoner of bourgeois origins, was at various times a courtier, diplomat, and civil servant. His poetry frequently (but not always unironically) reflects the views and values associated with the term "courtly." It is in some ways not easy to account for his decision to write in English, and it is not surprising that his earliest substantial poems, the *Book of the Duchess* (*c.* 1370) and the *House of Fame* (*c.* 1380), were heavily indebted to the fashionable French love-vision poetry of the time. Also of French origin was the octosyllabic couplet used in these poems. Chaucer's abandonment of this engaging but ultimately jejune metre in favour of a 10-syllable or iambic pentameter line was a portentous moment for English poetry. His mastery of it

was first revealed in stanzaic form, notably the seven-line stanza (rhyme royal) of the *Parlement of Foules* (*c.* 1382) and *Troilus and Criseyde* (*c.* 1385), and later was extended in the decasyllabic couplets of the prologue to the *Legend of Good Women* and large parts of *The Canterbury Tales.*

Though Chaucer wrote a number of moral and amatory lyrics, which were imitated by his 15th-century followers, his major achievements were in the field of narrative poetry. The early influence of French courtly love poetry (notably the *Roman de la Rose,* which he translated) gave way to an interest in Italian literature. Chaucer was acquainted with Dante's writings and took a story from Petrarch for the substance of his "Clerk's Tale." Two of his major poems, *Troilus and Criseyde* and "The Knight's Tale," were based, respectively, on the *Filostrato* and the *Teseida* of Boccaccio. The *Troilus,* Chaucer's single most ambitious poem, is a moving story of love gained and betrayed set against the background of the Trojan War. As well as being a poem of profound human sympathy and insight, it also has a marked philosophical dimension derived from Chaucer's reading of Boethius' *De consolatione philosophiae,* a work that he also translated in prose.

Chaucer's *Canterbury Tales*

His consummate skill in narrative art, however, was most fully displayed in *The Canterbury Tales* (*c.* 1387–1400), an unfinished series of stories purporting to be told by a group of pilgrims journeying from London to the shrine of St. Thomas Becket and back. The illusion that the individual pilgrims (rather than Chaucer himself) tell their tales gave him an unprecedented freedom of authorial stance, which enabled him to explore the rich fictive potentialities of a number of genres: pious legend (in "The Man of Law's Tale" and "The Prioress's Tale"), fabliaux ("The Shipman's Tale," "The Miller's Tale," and "The Reeve's Tale"), chivalric romance ("The Knight's Tale"), popular romance (parodied in Chaucer's "own" "Tale of Sir Thopas"), beast fable ("The Nun's Priest's Tale" and "The Manciple's Tale") and more—what Dryden later summed up as "God's plenty."

A recurrent concern in Chaucer's writings was the refined and sophisticated cultivation of love, commonly described by the modern expression "courtly love." A contemporary French term, *fine amour,* gives a more authentic description of the phenomenon; Chaucer's friend John Gower translated it as "fine loving" in his long poem *Confessio amantis* (begun *c.* 1386). The *Confessio* runs to some 33,000 lines in octosyllabic couplets and takes the form of a collection of exemplary tales placed within the framework of a lover's confession to a priest of Venus. Gower provides an interesting and sometimes refreshing contrast to Chaucer, in that the sober and earnest moral intent behind his writing is always clear, whereas Chaucer can be irritatingly noncommital and evasive. On the other hand, though Gower's verse is generally fluent and pleasing to read, it has a thin homogeneity of texture that cannot compare with the colour and range to be found in the language of his great contemporary. Gower was undoubtedly extremely learned by lay standards, and many classical myths (especially those deriving from Ovid's *Metamorphoses*) make the first of their numerous appearances in English literature in the *Confessio.* He was also deeply concerned with the moral and social condition of contemporary society, and he dealt with it in two weighty compositions in French and Latin, respectively: the *Mirour de l'omme* (*c.* 1374–78; "The Mirror of Man") and *Vox clamantis* (*c.* 1385).

Gower's *Confessio amantis*

Poetry after Chaucer and Gower. *Courtly poetry.* The numerous 15th-century followers of Chaucer continued to treat the conventional range of courtly and moralizing topics, but only rarely with the intelligence and stylistic accomplishment of their distinguished predecessors. The canon of Chaucer's works began to accumulate delightful but apocryphal trifles such as "The Flower and the Leaf" and "The Assembly of Ladies" (both *c.* 1475), the former, like a surprising quantity of 15th-century verse of this type, purportedly written by a woman. The stock figures of the ardent but endlessly frustrated lover and the irresistible but disdainful lady were cultivated as part of the "game of love" depicted in numerous courtly lyrics. Vernacular literacy spread rapidly among both lay men and women, the influence of French courtly love poetry remaining strong. Aristocratic and knightly versifiers such as Charles, duc d'Orléans (captured at Agincourt in 1415), his "jailer" William de la Pole, duke of Suffolk, and Sir Richard Ros (translator of Alain Chartier's influential *La Belle Dame sans merci*) were widely read and imitated among the gentry and in bourgeois circles well into the 16th century.

Both Chaucer and Gower had to some extent enjoyed royal and aristocratic patronage, and the active seeking of patronage became a pervasive feature of the 15th-century literary scene. Thomas Hoccleve, a minor civil servant who probably knew Chaucer and claimed to be his disciple, dedicated his *Regiment of Princes* (*c.* 1412), culled from an earlier work of the same name, to the future Henry V. Most of Hoccleve's compositions seem to have been written with an eye to patronage, and though they occasionally yield interesting and unexpected glimpses of his daily and private lives, they have little to recommend them as poetry. Hoccleve's aspiration to be Chaucer's successor was rapidly overshadowed, in sheer bulk if not necessarily in literary merit, by the formidable oeuvre of John Lydgate, a monk at the abbey of Bury St. Edmunds. Lydgate, too, was greatly stimulated at the prospects opened up by distinguished patronage, producing as a result a number of very long pieces that were greatly admired in their day. A staunch Lancastrian, Lydgate dedicated his *Troy Book* and *Life of Our Lady* to Henry V and his *Fall of Princes* (based ultimately on Boccaccio's *De casibus virorum illustrium*) to Humphrey Plantagenet, duke of Gloucester. He also essayed courtly verse in Chaucer's manner (*The Complaint of the Black Knight* and *The Temple of Glas*), but his imitation of the master's style was rarely successful. Both Lydgate and Hoccleve admired above all Chaucer's "eloquence," by which they meant mainly the Latinate elements in his diction. Their own painfully polysyllabic or "aureate" style unfortunately came to be widely imitated for more than a century. In sum, the major 15th-century English poets were generally undistinguished as successors of Chaucer, and for a significant but independent extension of his achievement one must look to the Scots *makaris* ("makers"), among whom were King James I of Scotland, Robert Henryson, and William Dunbar.

John Lydgate

Lydgate's following at court gave him a central place in 15th-century literary life, but the typical concerns shown by his verse do not distinguish it from a great body of religious, moral, historical, and didactic writing, much of it anonymous. A few identifiable provincial writers turn out to have had their own local patrons, often among the country gentry. East Anglia may be said to have produced a minor school in the works of John Capgrave, Osbern Bokenam, and John Metham, among others also active around the middle of the century. Some of the most moving and accomplished verse of the time is to be found in the anonymous lyrics and carols (songs with a refrain) on conventional subjects such as the transience of life, the coming of death, the sufferings of Christ, and other penitential themes. The author of some distinctive poems in this mode was John Audelay of Shropshire, whose style was heavily influenced by the alliterative movement. Literary devotion to the Virgin Mary was particularly prominent and at its best could produce masterpieces of artful simplicity, such as the justly famous "I sing of a maiden that is makeless."

Provincial writers

Popular and secular verse. The art that conceals art was also characteristic of the best popular and secular verse of the period, outside the courtly mode. Some of the shorter verse romances, usually in a form called tail rhyme, were far from negligible: *Ywain and Gawain* from the *Yvain* of Chrétien de Troyes; *Sir Launfal,* after Marie de France's *Lanval;* and *Sir Degrevant.* Humorous and lewd songs, versified tales, folk songs, ballads, and others form a lively but essentially subliterary body of compositions. Oral transmission was probably common, and the survival of much of what is extant is fortuitous. The Percy Folio manuscript, a 17th-century antiquarian collection of such material, may be a fair sampling of the repertoire of the late medieval itinerant entertainer. In addition to a number of more or less execrable popular romances of the

type satirized long before by Chaucer in "Sir Thopas," the Percy manuscript also contains a number of impressive ballads very much like those collected from oral sources in the 18th and 19th centuries. The extent of medieval origin of the poems collected in Francis J. Child's *English and Scottish Popular Ballads* (1882–98) is debatable. Several of the Robin Hood ballads undoubtedly were known in the 15th century, and the characteristic laconically repetitious and incremental style of the ballads is also to be seen in the enigmatic *Corpus Christi Carol,* preserved in an early 16th-century London grocer's commonplace book. In the same manuscript, but in a rather different vein, is *The Nut-Brown Maid,* an enchanting and expertly managed dialogue-poem on female constancy.

Political verse. A genre that does not fit easily into the categories already mentioned is political verse, of which a good deal was written in the 15th century. Much of it was avowedly and often crudely propagandist, especially during the Wars of the Roses, though a piece like the *Agincourt Carol* shows that it was already possible to strike the characteristically English note of insular patriotism soon after 1415. Of particular interest is the *Libel of English Policy* (*c.* 1436) on another typically English theme of a related kind: "Cherish merchandise, keep the admiralty,/ That we be masters of the narrow sea."

LATER MIDDLE ENGLISH PROSE

The continuity of a tradition in English prose writing, linking the later with the early Middle English period, is somewhat clearer than that to be detected in verse. The *Ancrene Wisse,* for example, continued to be copied and adapted to suit changing tastes and circumstances. But sudden and brilliant imaginative phenomena like the writings of Chaucer, Langland, and the author of *Sir Gawayne* are not to be found. Instead, there is a steady growth in the composition of religious prose of various kinds and the first appearance of secular prose in any quantity.

Religious prose. Of the first importance was the development of a sober, analytical, but nonetheless impressive kind of contemplative or mystical prose, represented by Walter Hilton's *Scale of Perfection* and the anonymous *Cloud of Unknowing.* The authors of these pieces certainly knew the more rugged and fervent writings of their earlier 14th-century predecessor Richard Rolle, and to some extent they reacted against what they saw as excesses in the style and content of his work. It is of particular interest to note that the mystical tradition was continued into the 15th century, though in very different ways, by two women writers, Julian of Norwich and Margery Kempe of King's Lynn. Julian, often regarded as the first English woman of letters, underwent a series of mystical experiences in 1373 about which she went on to write in her *Revelations of Divine Love,* one of the foremost works of English spirituality by the standards of any age. Rather different religious experiences went into the making of *The Book of Margery Kempe* (*c.* 1438), the extraordinary autobiographical record of a highly emotional bourgeoise, apparently dictated to a priest. The nature and status of its spiritual content remain controversial, but its often engaging colloquial style and vivid realization of the medieval scene are of abiding interest.

Mystical writings

Another important branch of the contemplative movement in prose involved the translation of continental Latin texts. A major example, and one of the best loved of all medieval English books in its time, was *The Mirror of the Blessed Life of Jesus Christ* (*c.* 1410), Nicholas Love's translation of the *Meditationes vitae Christi,* attributed to St. Bonaventure. Love's work was particularly valued by the church as an orthodox counterbalance to the heretical tendencies of the Lollards, who espoused the teachings of John Wycliffe and his circle. The Lollard movement generated a good deal of interesting and stylistically distinctive prose writing, though as the Lollards soon came under threat of death by burning, nearly all of it remains anonymous. A number of English works have been attributed to Wycliffe himself, and the first English translation of the Bible to Wycliffe's disciple John Purvey, but there are no firm grounds for these attributions. The Lollard Bible, which exists in a crude early form and in a more impressive later version (supposedly Purvey's work), was widely read in spite of being under doctrinal suspicion. It later influenced William Tyndale's translation of the New Testament, completed in 1525, and, through Tyndale, the Authorized Version (1611).

Secular prose. Secular compositions and translations in prose also came into prominence in the last quarter of the 14th century, though their stylistic accomplishment does not always match that of the religious tradition. Chaucer's "Tale of Melibeus" and his two astronomical translations, the *Treatise on the Astrolabe* and the *Equatorie of the Planetis,* were relatively modest endeavours beside the massive efforts of John of Trevisa, who translated from Latin both Ranulph Higden's universal history, *Polychronicon* (*c.* 1385–87), and Bartholomaeus Anglicus' encyclopaedia *De proprietatibus rerum* (1398). Judging by the number of surviving manuscripts, however, the most widely read secular prose work of the period is likely to have been *The Travels of Sir John Mandeville,* the supposed adventures of Sir John Mandeville, knight of St. Albans, on his journeys through Asia to the Orient. Though the work now is believed to be purely fictional, the exotic allure of the *Travels* and the occasionally arch style of their author were popular with the English reading public down to the 18th century.

Mandeville's *Travels*

The 15th century saw the consolidation of English prose as a respectable medium for serious writings of various kinds. The anonymous *Brut* chronicle survives in more manuscripts than any other medieval English work and was instrumental in fostering a new sense of national identity. John Capgrave's *Chronicle of England* (*c.* 1462) and Sir John Fortescue's *On the Governance of England* (*c.* 1470) were part of the same trend. At its best, the style of such works could be vigorous and straightforward, close to the language of everyday speech, like that found in the chance survivals of private letters of the period. Best known and most numerous among letters are those of the Paston family of Norfolk, but significant collections were also left by the Celys of London and the Stonors of Oxfordshire. More eccentric prose stylists of the period were the religious controversialist Reginald Pecock and John Skelton, whose "aureate" translation of the *Bibliotheca historica* of Diodorus Siculus stands in marked contrast to the demotic exuberance of his verse.

The crowning achievement of later Middle English prose writing was Sir Thomas Malory's cycle of Arthurian legends, which was given the title *Le Morte Darthur* by William Caxton when he printed his edition in 1485. There is still uncertainty as to the identity of Malory, who described himself as a "knight-prisoner." The characteristic mixture of chivalric nostalgia and tragic feeling with which he imbued his book gave fresh inspiration to the tradition of writing on Arthurian themes. The nature of Malory's artistry eludes easy definition, and the degree to which the effects he achieved were a matter of conscious contrivance on his part is debatable. Much of the *Morte Darthur* was translated from prolix French prose romances, and Malory evidently selected and condensed his material with instinctive mastery as he went along. At the same time he cast narrative and dialogue in the cadences of a virile and natural English prose that admirably matched the nobility of both the characters and the theme.

Malory's Arthurian cycle

MIDDLE ENGLISH DRAMA

Because the manuscripts of medieval English plays were usually ephemeral performance scripts rather than reading matter, very few examples have survived from what once must have been a very large dramatic literature. What little survives from before the 15th century includes some bilingual fragments, indicating that the same play might have been given in English or Anglo-Norman, according to the composition of the audience. From the late 14th century onward two main dramatic genres are discernible, the mystery or Corpus Christi cycles and the morality plays. The mystery plays were long cyclic dramas of the Creation, Fall, and Redemption of mankind, based mostly on biblical narratives. They usually included a selection of Old Testament episodes (such as the stories of Cain and Abel and Abraham and Isaac) but concentrated mainly on

Mystery and morality plays

the life and Passion of Jesus Christ. They always ended with the Last Judgment. The cycles were generally financed and performed by the craft guilds and staged on wagons in the streets and squares of the towns. Texts of the cycles staged at York, Chester, Wakefield, and at an unstated location in East Anglia have survived, together with fragments from Coventry, Newcastle, and Norwich. Their literary quality is uneven, but the York cycle (probably the oldest) has a most impressively realized version of Christ's Passion by a dramatist influenced by the alliterative style in verse. Wakefield has several particularly brilliant plays, attributed to the anonymous Wakefield Master, and his *Second Shepherds' Play* is one of the masterpieces of medieval English literature. The morality plays were allegorical dramas depicting the progress of a single character, representing the whole of mankind, from the cradle to the grave and sometimes beyond. The other dramatis personae might include God and the Devil but usually consisted of personified abstractions, such as the Vices and Virtues, Death, Penance, Mercy, and so forth. An interesting and varied collection of the moralities is known as the Macro Plays (*The Castle of Perseverance, Wisdom, Mankind*), but the single most impressive piece is undoubtedly *Everyman,* a superb English rendering of a Dutch play on the subject of the coming of death. Both the mystery and morality plays have been frequently revived and performed in the 20th century.

THE TRANSITION FROM MEDIEVAL TO RENAISSANCE

Transition from medieval to Renaissance

The 15th century was a major period of growth in lay literacy, a process powerfully expedited by the introduction into England of printing by William Caxton in 1476. Caxton's Malory (1485) was published in the same year that Henry Tudor acceded to the throne as Henry VII, and the period from this time to the mid-16th century has been called the transition from medieval to Renaissance in English literature. A typical figure was the translator Alexander Barclay. His *Eclogues* (*c.* 1515), drawn from 15th-century Italian humanist sources, was an early essay in the fashionable Renaissance genre of pastoral, while his rendering of Sebastian Brant's *Narrenschiff* as *The Ship of Fools* (1509) is a thoroughly medieval satire on contemporary folly and corruption. *The Passetyme of Pleasure* (1506) by Stephen Hawes, ostensibly an allegorical romance in Lydgate's manner, unexpectedly adumbrates the great Tudor theme of academic cultivation as a necessary accomplishment of the courtly knight or gentleman.

The themes of education and good government predominate in the new humanist writing of the 16th century, both in discursive prose (such as Sir Thomas Elyot's *Boke Named the Governour* and Roger Ascham's *Toxophilus* and *Scholemaster*) and in the drama (the plays of Henry Medwall and Richard Rastall). The preeminent work of English humanism, Sir Thomas More's *Utopia* (1516), was composed in Latin and appeared in an English translation in 1551. Undoubtedly the most distinctive voice in the poetry of the time was that of John Skelton, tutor to Henry VII's sons and author of an extraordinary range of writing, often in an equally extraordinary style. His works include a long play, *Magnyfycence,* like his *Bowge of Courte* an allegorical satire on court intrigue; intemperate satirical invectives, such as *Collyn Clout* and *Why Come Ye Nat to Courte?* (both 1522); and unusual reflexive essays on the role of the poet and poetry, in *Speke, Parrot* (written 1521) and *The Garland of Laurel* (1523). The first half of the 16th century was also a notable period for courtly lyric verse in the stricter sense of poems with musical settings, such as those found in the Devonshire manuscript. This is very much the literary milieu of the "courtly makers" Sir Thomas Wyatt and Henry Howard, earl of Surrey, but though the courtly context of much of their writing is of medieval origin, their most distinctive achievements look to the future. Poems like Wyatt's "They flee from me" and "Whoso list to hunt" vibrate with personal feeling at odds with the medieval convention of anonymity, while Surrey's translations from the *Aeneid* introduce blank verse (unrhymed iambic pentameter) into English for the first time, providing an essential foundation for the achievements of Shakespeare and Milton. (Ri.B.)

The Renaissance period: 1550–1660

LITERATURE AND THE AGE

In a tradition of literature remarkable for its exacting and brilliant achievements, the Elizabethan and early Stuart periods have been said to represent the most brilliant century of all. (The reign of Elizabeth I began in 1558 and ended with her death in 1603; she was succeeded by the Stuart king James VI of Scotland, who took the title James I of England as well. English literature of his reign as James I, from 1603 to 1625, is properly called Jacobean.) These years produced a gallery of authors of genius, some of whom have never been surpassed, and conferred on scores of lesser talents the enviable ability to write with fluency, imagination, and verve. From one point of view, this sudden renaissance looks radiant, confident, heroic—and belated, but all the more dazzling for its belatedness. Yet from another point of view, this was a time of unusually traumatic strain, in which English society underwent massive disruptions that transformed it on every front and decisively affected the life of every individual. In the brief, intense moment in which England assimilated the European Renaissance, the circumstances that made the assimilation possible were already disintegrating and calling into question the newly won certainties, as well as the older truths that they were dislodging. This doubleness, of new possibilities and new doubts simultaneously apprehended, gives the literature its unrivaled intensity.

Social conditions. In this period England's population doubled; prices rocketed, rents followed, old social loyalties dissolved, and new industrial, agricultural, and commercial veins were first tapped. Real wages hit an all-time low in the 1620s, and social relations were plunged into a state of unprecedented fluidity from which the merchant and ambitious lesser gentleman profited at the expense of the aristocrat and labourer, as satires and comedies current from the 1590s complain. Behind the Elizabethan vogue for pastoral poetry lies the fact of the prosperity of the enclosing sheep farmer, who aggressively sought to increase pasture at the expense of the peasantry. Tudor platitudes about order and degree could neither combat nor survive the challenge posed to rank by these arrivistes. The position of the crown, politically dominant yet financially insecure, had always been potentially unstable, and when Charles I lost the confidence of his greater subjects in the 1640s his authority crumbled. Meanwhile, the huge body of poor fell ever further behind the rich; the pamphlets of Thomas Harman (1566) and Robert Greene (1591–92), and Shakespeare's *King Lear* (1605), provide glimpses of a horrific world of vagabondage and crime, the Elizabethans' biggest, unsolvable social problem.

Intellectual and religious revolution. The barely disguised social ferment was accompanied by an intellectual revolution, as the medieval synthesis collapsed before the new science, new religion, and new humanism. While modern mechanical technologies were pressed into service by the Stuarts to create the scenic wonders of the court masque, the discoveries of astronomers and explorers were redrawing the cosmos in a way that was profoundly disturbing:

> And freely men confess that this world's spent,
> When in the planets, and the firmament
> They seek so many new . . .
>
> (John Donne, *The First Anniversary,* 1611)

The majority of people were more immediately affected by the religious revolutions of the 16th century. The man in early adulthood at the accession of Elizabeth in 1558 would, by her death in 1603, have been vouchsafed an unusually disillusioning insight into the duty owed by private conscience to the needs of the state. The Tudor church was an instrument of social and political coercion, yet the mid-century controversies over the faith had already wrecked any easy confidence in the authority of doctrines and forms and had taught men to question carefully the rationale of their own beliefs (as Donne does in his third *Satire, c.* 1596). The Elizabethan ecclesiastical compromise was the object of continual criticism, both from radicals within (who desired progressive reforms, such as the abolition of bishops) and from papists without (who

desired the return of England to the Roman Catholic fold), but the incipient liberalism of individuals like John Milton and William Chillingworth was held in check by the majority's unwillingness to tolerate a plurality of religions in a supposedly unitary state. Nor was the Calvinist orthodoxy that cradled most English writers comforting, for it told them that they were corrupt, unfree, unable to earn their own salvations, and subject to heavenly judgments that were arbitrary and absolute. It deeply informs the world of the Jacobean tragedies, whose heroes are not masters of their fates but victims of divine purposes that are terrifying yet inscrutable.

The race for cultural development. The third complicating factor was the race to catch up with continental developments in arts and philosophy. The Tudors badly needed to create a class of educated diplomats, statesmen, and officials and to dignify their court by making it a fount of cultural as well as political patronage. The new learning, widely disseminated through the Erasmian educational programs of such men as John Colet and Sir Thomas Elyot, proposed to use a systematic schooling in Latin authors and some Greek to encourage in the social elites a flexibility of mind and civilized serviceableness by which enlightened princely government could walk hand in hand with responsible scholarship. Humanism fostered an intimate familiarity with the classics that was a powerful incentive for the creation of an English literature of answerable dignity. It fostered as well a practical, secular piety that left its impress everywhere on Elizabethan writing. Humanism's effect, however, was modified by the simultaneous impact of the flourishing continental cultures, particularly the Italian. Repeatedly, crucial innovations in English letters developed resources originating from Italy, such as the sonnet of Petrarch, the epic of Ariosto, the pastoral of Sannazzaro, the canzone, and blank verse, and values imported with these forms were in competition with the humanists' ethical preoccupations. Social ideals of wit, many-sidedness, and *sprezzatura* (accomplishment mixed with unaffectedness) were imbibed from Baldassare Castiglione's *Il cortegiano,* translated as *The Courtyer* by Sir Thomas Hoby in 1561, and Elizabethan court poetry is steeped in Castiglione's aristocratic Neoplatonism, his notions of universal proportion, and the love of beauty as the path to virtue. Equally significant was the welcome afforded to Niccolò Machiavelli, whose lessons were vilified publicly and absorbed in private. *The Prince,* written in 1513, was unavailable in English until 1640, but as early as the 1580s Gabriel Harvey, a friend of the poet Edmund Spenser, can be found enthusiastically hailing its author as the apostle of modern pragmatism. "We are much beholden to Machiavel and others," said Bacon, "that write what men do, and not what they ought to do."

The mark of humanism

So the literary revival occurred in a society deeply torn and rife with tensions, uncertainties, and competing versions of order and authority, religion and status, sex and the self. The Elizabethan compromise was exactly that; the Tudor pretense that all the nation thought the same disguised the actual fragmentation of the old consensus under the strain of change. The new scientific knowledge proved both man's littleness and his power to command nature; against the Calvinist idea of man's helplessness pulled the humanist faith in his dignity, especially that conviction, derived from the reading of Seneca and so characteristic of the period, of man's constancy and fortitude, his heroic and almost divine capacity for self-determination. It was still possible for Elizabeth to hold these divergent tendencies together in a single, heterogeneous culture, but under her successors they would eventually fly apart. The philosophers speaking for the new century would be Francis Bacon, who argued for the gradual advancement of science through patient accumulation of experiments, and the skeptic Michel de Montaigne (his *Essais* translated from the French by John Florio, 1603), who denied that it was possible to formulate any general principles of knowledge.

Cutting across all of these was the persistence of popular habits of thought and expression. Both humanism and puritanism set themselves against vulgar ignorance and folk tradition, but, fortunately, neither could remain aloof for long from the robustness of popular taste. Sir Philip Sidney, in England's first neoclassical literary treatise, *The Defence of Poesie* (written *c.* 1578–1583, published 1595), candidly admitted that "the old song of Percy and Douglas" would move his heart "more than with a trumpet," and his *Arcadia* is a representative instance of the continual, fruitful cross-fertilization of genres in this period—the contamination of aristocratic pastoral with popular tale, the lyric with the ballad, comedy with romance, tragedy with satire, and poetry with prose. The language, too, was undergoing a rapid expansion that all classes contributed to and benefited from, sophisticated literature borrowing without shame the idioms of colloquial speech. Macbeth's allusion to heaven peeping "through the blanket of the dark" only became a problem in an age when tragic dignity implied politeness, when it was below the dignity of a tragic hero to mention so lowly an object as a blanket. The Elizabethans' ability to address themselves to several audiences simultaneously and to bring into relation opposed experiences, emphases, and worldviews invested their writing with complexity and power.

Influence of popular taste

ELIZABETHAN POETRY AND PROSE

English poetry and prose burst into sudden glory in the late 1570s. A decisive shift of taste toward a fluent artistry self-consciously displaying its own grace and sophistication was announced in the works of Spenser and Sidney. It was accompanied by an upsurge in literary production that came to fruition in the 1590s and 1600s, two decades of astonishing productivity by writers of every persuasion and calibre.

The groundwork was laid in the 30 years from 1550, a period of slowly increasing confidence in the literary competence of the language and tremendous advances in education, which for the first time produced a substantial English readership, keen for literature and possessing cultivated tastes. This development was underpinned by the technological maturity and accelerating output (mainly in pious or technical subjects) of Elizabethan printing. The Stationers' Company, which controlled the publication of books, was incorporated in 1557, and Richard Tottel's *Miscellany* (1557) revolutionized the relationship of poet and audience by making publicly available lyric poetry, which hitherto had circulated only among a courtly coterie. Edmund Spenser was the first considerable English poet deliberately to use print for the advertisement of his talents.

Tottel's *Miscellany*

Development of the English language. The prevailing opinion of the language's inadequacy, its lack of "terms" and innate inferiority to the eloquent classical tongues, was combated in the work of the humanists Thomas Wilson, Roger Ascham, and Sir John Cheke, whose treatises on rhetoric, education, and even archery argued in favour of an unaffected vernacular prose and a judicious attitude toward linguistic borrowings. Their stylistic ideals are attractively embodied in Ascham's educational tract *The Scholemaster* (1570), and their tonic effect on that particularly Elizabethan art, translation, can be felt in the earliest important examples, Sir Thomas Hoby's Castiglione (1561) and Sir Thomas North's Plutarch (1579). A further stimulus was the religious upheaval that took place in the middle of the century. The desire of Reformers to address as comprehensive an audience as possible—the bishop and the boy who follows the plough, as Tyndale put it—produced the first true classics of English prose: the reformed Anglican Book of Common Prayer (1549, 1552, 1559); John Foxe's *Actes and Monuments* (1563), which celebrates the martyrs, great and small, of English Protestantism; and the various English versions of Scripture, from William Tyndale's (1525), Miles Coverdale's (1535), and the Geneva Bible (1560) to the syncretic Authorized Version (1611). The latter's combination of grandeur and plainness is justly celebrated, even if it represents an idiom never spoken in heaven or on earth. Nationalism inspired by the Reformation motivated the historical chronicles of the capable and stylish Edward Hall (1548), who bequeathed to Shakespeare the tendentious Tudor interpretation of the 15th century, and of the rather less capable Raphael Holinshed (1577). John

Ponet's remarkable *Short Treatise of Politic Power* (1556) is a vigorous polemic against Mary Tudor, whom he saw as a papist tyrant.

In verse, Tottel's much reprinted *Miscellany* generated a series of imitations and, by popularizing the lyrics of Wyatt and Surrey, carried into the 1570s the tastes of the early Tudor court. The newer poets collected by Tottel and other anthologists include Nicholas Grimald, Richard Edwardes, George Turberville, Barnabe Googe, George Gascoigne, Sir John Harington, and many others, of whom Gascoigne is the most considerable. The modern preference for the ornamental manner of the next generation has eclipsed these poets, who continued the tradition of plain, weighty verse, addressing themselves to ethical and didactic themes and favouring the meditative lyric, satire, and epigram. But their taste for economy, restraint, and aphoristic density was, in the verse of Ben Jonson and Donne, to outlive the cult of elegance. The period's major project was *A Mirror for Magistrates* (1559; enlarged editions 1563, 1578, 1587), a collection of verse laments, by several hands, purporting to be spoken by participants in the Wars of the Roses and preaching the Tudor doctrine of obedience. The quality is uneven, but Thomas Sackville's "Induction" and Thomas Churchyard's *Legend of Shore's Wife* are distinguished, and the intermingling of history, tragedy, and political morality was to be influential on the drama.

Sidney and Spenser. With the work of Sidney and Spenser, Tottel's contributors suddenly began to look old-fashioned. Sir Philip Sidney epitomized the new Renaissance "universal man": a courtier, diplomat, soldier, and poet whose *Defence of Poesie* included the first considered account of the state of English letters. Sidney's treatise defends literature on the ground of its unique power to teach, but his real emphasis is on its delight, its ability to depict the world not as it is but as it ought to be. This quality of "forcefulness or *energia*" he himself demonstrated in his sonnet sequence of unrequited desire, *Astrophel and Stella* (written *c.* 1582, published 1591). His *Arcadia,* in its first version (written *c.* 1577–80), is a pastoral romance in which courtiers disguised as Amazons and shepherds make love and sing delicate experimental verses. The revised version (written *c.* 1580–84, published 1590), vastly expanded but abandoned in mid-sentence, added sprawling plots of heroism in love and war, philosophical and political discourses, and set pieces of aristocratic etiquette. Sidney was a dazzling and assured innovator whose pioneering of new forms and stylistic melody was seminal for his generation. His public fame was as an aristocratic champion of an aggressively Protestant foreign policy, but Elizabeth had no time for idealistic warmongering, and thus his fictions abound with situations of inhibition and withheld satisfactions—unresolved conflicts of desire against restraint, heroism against patience, rebellion against submission—that mirror his own position as an unsuccessful courtier.

Sidney's *Defence of Poesie*

Protestantism also loomed large in the life of Edmund Spenser. He enjoyed the patronage of the Earl of Leicester, who sought to advance militant Protestantism at court, and his poetic manifesto, *The Shepheardes Calender* (1579), covertly praised Archbishop Edmund Grindal, who had been suspended by Elizabeth for his Puritan sympathies. Spenser's masterpiece, *The Faerie Queene* (1590–1609), is an epic of Protestant nationalism in which the villains are infidels or papists, the hero is King Arthur, and the central value is married chastity.

Spenser's *Faerie Queene*

Spenser was one of the humanistically trained breed of public servants, and the *Calender,* an expertly crafted collection of pastoral eclogues, both advertised his talents and announced his epic ambitions, the exquisite lyric gift that it reveals being voiced again in the marriage poems *Epithalamion* (1595) and *Prothalamion* (1596). With *The Faerie Queene* he achieved the central poem of the Elizabethan period. Its form fuses the medieval allegory with the Italian romantic epic; its purpose was "to fashion a gentleman or noble person in virtuous and gentle discipline." The plan was for 12 books (six were completed), focusing on 12 virtues exemplified in the quests of 12 knights from the court of Gloriana, the Faerie Queene, a symbol for Elizabeth herself. Arthur, in quest of Gloriana's love, would appear in each book and come to exemplify Magnificence, the complete man. Spenser took the decorative chivalry of the Elizabethan court festivals and reworked it through a constantly shifting veil of allegory, so that the knight's adventures and loves build into a complex, multileveled portrayal of the moral life. The verse, a spacious and slow-moving nine-lined stanza, and archaic language frequently rise to an unrivaled sensuousness.

The Faerie Queene was a public poem, addressed to the Queen, and politically it echoed the hopes of the Leicester circle for government motivated by godliness and militancy. Spenser's increasing disillusion with the court and with the active life, a disillusion noticeable in the later books and in his bitter satire *Colin Clouts Come Home Againe* of 1591, voiced the fading of these expectations in the last decade of Elizabeth's reign, the beginning of that remarkable failure of political and cultural confidence in the monarchy. In the "Mutabilitie Cantos," melancholy fragments of a projected seventh book, Spenser turned away from the public world altogether, toward the ambiguous consolations of eternity.

The lessons taught by Sidney and Spenser in the cultivation of melodic smoothness and graceful refinement appear to good effect in the subsequent virtuoso outpouring of lyrics and sonnets. These are among the most engaging achievements of the age, though the outpouring was itself partly a product of frustration, as a generation trained to expect office or preferment but faced with courtly parsimony channeled its energies in new directions in search of patronage. For Sidney's fellow courtiers, pastoral and love lyric were also a means of obliquely expressing one's relationship with the Queen, of advancing a proposal or an appeal.

Elizabethan lyric. Virtually every Elizabethan poet tried his hand at the lyric; few, if any, failed to write one that is not still anthologized today. The fashion for interspersing prose fiction with lyric interludes, begun in the *Arcadia,* was continued by Robert Greene and Thomas Lodge (notably in the latter's *Rosalynde,* 1590, the source for Shakespeare's *As You Like It*), and in the theatres plays of every kind were diversified by songs both popular and courtly. Fine examples are in the plays of John Lyly, George Peele, Thomas Nashe, Ben Jonson, and Thomas Dekker (though all, of course, are outshone by Shakespeare's). The most important influence, though, was the outstanding richness of late Tudor music, in both the native tradition of expressive lute song, represented by John Dowland, and the complex Italianate madrigal newly imported by William Byrd and Thomas Morley. The foremost talent among lyricists, Thomas Campion, was composer as well as poet; his songs (four *Bookes of Ayres,* 1601–17) are unsurpassed for their clarity, harmoniousness, and rhythmic subtlety. Even the work of a lesser talent, however, such as Nicholas Breton, is remarkable for the suggestion of depth and poise in the slightest performances; the smoothness and apparent spontaneity of Elizabethan lyric conceals a consciously ordered and laboured artifice, attentive to decorum and rhetorical fitness. These are not personal but public pieces, intended for singing and governed by a Neoplatonic aesthetic in which delight is a means of addressing the moral sense, harmonizing the auditor's mind and attuning it to the discipline of reason and virtue. This necessitates a deliberate narrowing of scope—to the readily comprehensible situations of pastoral or Petrarchan hope and despair—and makes for a certain uniformity of effect, albeit an agreeable one. The lesser talents are well displayed in the miscellanies *The Phoenix Nest* (1593), *Englands Helicon* (1600), and *A Poetical Rhapsody* (1602).

Campion's songs

The sonnet sequence. The publication of Sidney's *Astrophel and Stella* in 1591 generated an equally extraordinary vogue for the sonnet sequence, Sidney's principal imitators being Samuel Daniel, Michael Drayton, Fulke Greville, Spenser, and Shakespeare, and his lesser, Henry Constable, Barnabe Barnes, Giles Fletcher, Thomas Lodge, Richard Barnfield, and many more. *Astrophel* had re-created the Petrarchan world of proud beauty and despairing lover in a single, brilliant stroke, though in English hands the preferred division of the sonnet into three quatrains and a

couplet gave Petrarch's contemplative form a more forensic turn, investing it with an argumentative terseness and epigrammatic sting. Within the common ground shared by the sequences there is much diversity. Only Sidney's endeavours to tell a story, the others being more loosely organized as variations focusing on a central (usually fictional) relationship. Daniel's *Delia* (1592) is eloquent and elegant, dignified and high-minded; Drayton's *Ideas Mirrour* (1594; much revised by 1619) rises to a strongly imagined, passionate intensity; Spenser's *Amoretti* (1595) celebrates, eccentrically, fulfilled sexual love achieved within marriage. Shakespeare's sonnets (published 1609) present a different world altogether, the conventions upside down, the lady no beauty but dark and treacherous, the loved one genuinely beyond considerations of sexual possession because he is a boy. The sonnet tended to gravitate toward correctness or politeness, and for most readers its chief pleasure must have been rhetorical, in its forceful pleading and consciously exhibited artifice, but under the pressure of Shakespeare's urgent metaphysical concerns, dramatic toughness, and shifting and highly charged ironies, the form's conventional limits were exploded.

Other poetic styles. Sonnet and lyric represent one tradition of verse within the period, that most conventionally delineated as Elizabethan, but the picture is complicated by the coexistence of other poetic styles in which ornament was distrusted or turned to different purposes; the sonnet was even parodied by Sir John Davies in his *Gulling Sonnets* (*c.* 1594) and by the Jesuit poet Robert Southwell. A particular stimulus to experiment was the variety of new possibilities made available by verse translation, from Richard Stanyhurst's extraordinary *Aeneid* (1582), in quantitative hexameter and littered with obscure or invented diction, and Sir John Harington's version of Ariosto's *Orlando furioso* (1591), with its Byronic ease and narrative fluency, to Christopher Marlowe's blank verse rendering of *Lucan's First Book* (published 1600), probably the finest Elizabethan translation.

Epyllion

The genre to benefit most from translation was the epyllion, or little epic. This short narrative in verse was usually on a mythological subject, taking most of its material from Ovid, either his *Metamorphoses* (English version by Arthur Golding, 1565–67) or his *Heroides* (English version by Turberville, 1567). This form flourished from Thomas Lodge's *Scillaes Metamorphosis* (1589) to Francis Beaumont's *Salmacis and Hermaphroditus* (1602) and is best represented by Marlowe's *Hero and Leander* (published 1598) and Shakespeare's *Venus and Adonis* (1593). Ovid's reputation as an esoteric philosopher left its mark on George Chapman's *Ovid's Banquet of Sense* (1595) and Drayton's *Endimion and Phoebe* (1595), in which the love of mortal for goddess becomes a parable of wisdom. But his real attraction was as an authority on the erotic, and most epyllia treat physical love with sophistication and sympathy, unrelieved by the gloss of allegory—a tendency culminating in John Marston's *The Metamorphosis of Pigmalion's Image* (1598), a poem that has shocked tender sensibilities. Inevitably, the shift of attitude had an effect on style: for Marlowe the experience of translating (inaccurately) Ovid's *Amores* meant a gain for *Hero and Leander* in terms of urbanity and, more important, wit.

With the epyllion comes a hint of the tastes of the following reign, and a similar shift of taste can be felt among those poets of the 1590s who began to modify the ornamental style in the direction of native plainness or classical restraint. An astute courtier like Sir John Davies might, in his *Orchestra* (1596) and *Hymns of Astraea* (1599), write confident panegyrics to the aging Elizabeth, but in Sir Walter Raleigh's "Eleventh Book of the Ocean to Cynthia," a kind of broken pastoral eclogue, praise of the Queen is undermined by an obscure but eloquent sense of hopelessness and disillusionment. For Raleigh the complimental manner seems to be disintegrating under the weight of disgrace and isolation at court; his scattered lyrics, notably that contemptuous dismissal of the court, "The Lie," often draw their resonance from the resources of the plain style. Another courtier whose writing suggests similar pressures is Fulke Greville, Lord Brooke. Greville's *Caelica* (published 1633) begins as a conventional sonnet sequence but gradually abandons Neoplatonism for pessimistic reflections on religion and politics. Other works in his sinewy and demanding verse include philosophical treatises and unperformed melodramas (*Alaham* and *Mustapha*) that have a sombre Calvinist tone, presenting man as a vulnerable creature inhabiting a world of unresolved contradictions:

> Oh wearisome condition of humanity!
> Born under one law, to another bound;
> Vainly begot, and yet forbidden vanity,
> Created sick, commanded to be sound.
>
> (*Mustapha,* chorus)

Greville was a friend of the Earl of Essex, whose revolt against Elizabeth ended in 1601 on the scaffold, and other poets on the edge of the Essex circle fueled the taste for aristocratic heroism and individualist ethics. George Chapman's masterpiece, his translation of Homer (1598), is dedicated to Essex, and his original poems are intellectual and recondite, often deliberately cultivating obscurities; his abstruseness is a means of restricting his audience to a worthy, understanding elite. Samuel Daniel, in his verse *Epistles* (1603) written to various noblemen, strikes a mean between plainness and compliment; his *Musophilus* (1599), dedicated to Greville, defends the worth of poetry but says there are too many frivolous wits writing. The cast of Daniel's mind is stoical, and his language is classically precise. His major project was a verse history of *The Civil Wars between the Two Houses of Lancaster and York* (1595–1609), and versified history is also strongly represented in the *Legends* (1593–1607), *Barons' Wars* (1596, 1603), and *Englands Heroicall Epistles* (1597) of Michael Drayton.

Satire

The form really to set its face against Elizabethan politeness was the satire. Satire was related to the complaint, of which there were notable examples by Daniel (*The Complaint of Rosamond,* 1592) and Shakespeare (*The Rape of Lucrece,* 1594), and these are dignified and tragic laments in supple verse, but the Elizabethans mistakenly held the term satire to derive from the Greek *satyros,* a satyr, and so set out to match their manner to their matter and make their verses snarl. In the works of the principal satirists, John Donne (five satires, 1593–98), Joseph Hall (*Virgidemiarum,* 1597–98), and John Marston (*Certaine Satyres* and *The Scourge of Villainy,* 1598), the denunciation of vice and folly repeatedly tips into invective, raillery, and sheer abuse. The versification of Donne's satires is frequently so rough as barely to be verse at all; Hall apologized for not being harsh enough, and Marston was himself pilloried in Ben Jonson's play *Poetaster* (1601) for using ridiculously difficult language. "Vex all the world," wrote Marston to himself, "so that thyself be pleased." The satirists popularized a new persona, that of the malcontent who denounces his society not from above but from within, and their continuing attraction resides in their self-contradictory delight in the world they profess to abhor and their evident fascination with the minutiae of life in court and city. They were enthusiastically followed by Everard Guilpin, Samuel Rowlands, Thomas Middleton, and Cyril Tourneur, and so scandalous was the flood of satires that in 1599 their printing was banned. Thereafter the form survived in Jonson's classically balanced epigrams and poems of the good life, but its more immediate impact was on the drama, in helping to create the vigorously skeptical voices that people *The Revenger's Tragedy* and *Hamlet.*

Prose styles. Description of the development of Elizabethan prose begins with the 1570s. Prose was easily the principal medium in the Elizabethan period, and, despite the mid-century uncertainties over the language's weaknesses and strengths—whether coined and imported words should be admitted; whether the structural modeling of English prose on Latin writing was beneficial or, as Bacon would complain, a pursuit of "choiceness of phrase" at the expense of "soundness of argument"—the general attainment of prose writing was uniformly high, as is often manifested in contexts not conventionally imaginative or "literary," such as tracts, pamphlets, and treatises. The obvious instance of such casual success is Richard Hakluyt's *Principall Navigations, Voiages, and*

Discoveries of the English Nation (1589; expanded 1598–1600), a massive collection of travelers' tales, of which some are highly accomplished narratives. William Harrison's gossipy, entertaining *Description of England* (1577), Philip Stubbes's excitable and humane social critique *The Anatomy of Abuses* (1583), Reginald Scot's anecdotal *Discovery of Witchcraft* (1584), and John Stow's invaluable *Survey of London* (1598) also deserve passing mention. William Kempe's account of his morris dance from London to Norwich, *Kempe's Nine Days' Wonder* (1600), has great charm.

Early prose fiction

The writers listed above all use an unpretentious style, enlivened with a vivid vocabulary; the early prose fiction, on the other hand, delights in ingenious formal embellishment at the expense of narrative economy. This runs up against preferences ingrained in the modern reader by the novel, but Elizabethan fiction is not at all novelistic and finds room for debate, song, and the conscious elaboration of style. The unique exception is George Gascoigne's "Adventures of Master F. J." (1573), a tale of thwarted love set in an English great house, which is the first success in English imaginative prose. Gascoigne's story has a surprising authenticity and almost psychological realism (it may be autobiographical), but even so it is heavily imbued with the influence of Castiglione.

The existence of an audience for polite fiction was signaled in the collections of stories imported from France and Italy by William Painter (1566), Geoffrey Fenton (1577), and George Pettie (1576). Pettie, who claimed not to care "to displease twenty men to please one woman," believed his readership was substantially female. There were later collections by Barnaby Rich (1581) and George Whetstone (1583); historically, their importance was as sources of plots for many Elizabethan plays. The direction fiction was to take was established by John Lyly's *Euphues: The Anatomy of Wit* (1578), which, with its sequel *Euphues and His England* (1580), set a fashion for an extreme rhetorical mannerism that came to be known as "euphuism." The priggish plot of *Euphues*—a rake's fall from virtue and his recovery—is but an excuse for a series of debates, letters, and speechifyings, thick with assonance, antithesis, parallelism, and balance and displaying a pseudoscientific learning. Lyly's style was to be successful on the stage, but in fiction its density and monotony are wearying. The other major prose work of the 1570s, Sidney's *Arcadia,* is no less rhetorical (Abraham Fraunce illustrated his handbook of style *The Arcadian Rhetoric,* 1588, almost entirely with examples from the *Arcadia*), but with Sidney rhetoric is in the service of psychological insight and an exciting plot. Dozens of imitations of *Arcadia* and *Euphues* followed from the pens of Robert Greene, Thomas Lodge, Anthony Munday, Emanuel Forde, and others; none has much distinction.

Prose was to be decisively transformed through its involvement in the bitter and learned controversies of the 1570s and '80s over the reform of the English Church and the problems the controversies raised in matters of authority, obedience, and conscience. The fragile ecclesiastical compromise threatened to collapse under the demands made by Elizabeth's more godly subjects for further reformation, and its defense culminated in Richard Hooker's *Of the Laws of Ecclesiastical Polity* (eight books, 1593–1662), the first English classic of serious prose. Hooker's

Writings on religious issues

is a monumental work, structured in massive and complex paragraphs brilliantly recreating the orotund style of Cicero. His air of maturity and detachment has recommended him to modern tastes, but no more than his opponents was he above the cut and thrust of controversy. On the contrary, his magisterial rhetoric was designed all the more effectively to fix blame onto his enemies, and even his account (in books VI–VIII) of the relationship of church and state was deemed too sensitive for publication in the 1590s.

More decisive for English fiction was the appearance of the "Martin Marprelate" tracts of 1588–90. These seven pamphlets argued the Puritan case but with an unpuritanical scurrility and created great scandal by hurling invective and abuse at Elizabeth's bishops with comical gusto. The bishops employed Lyly and Thomas Nashe to reply to Marprelate, and the consequence may be read in Nashe's prose satires of the following decade, especially *Piers Penniless His Supplication to the Devil* (1592), *The Unfortunate Traveller* (1594), and *Lenten Stuffe* (1599), the latter a mock encomium on red herring. Nashe's "extemporal vein" makes fullest use of the flexibility of colloquial speech and delights in nonsense, redundancy, and disconcerting shifts of tone, which demand an answering agility from the reader. His language is probably the most profusely inventive of all Elizabethan writers', and he even makes the low-life pamphlets of Robert Greene (1591–92), with their sensational tales from the underworld, look conventional. His only rival is Thomas Deloney, whose *Jack of Newbury* (1597), *The Gentle Craft* (1597–98), and *Thomas of Reading* (1600) are enduringly attractive for their depiction of the lives of ordinary citizens, interspersed with elements of romance, jest book, and folktale. Deloney's entirely convincing dialogue indicates how important for the development of a flexible prose must have been the example of a flourishing theatre in Elizabethan London. In this respect, as in so many others, the role of the drama was crucial.

ELIZABETHAN AND EARLY STUART DRAMA

Theatre and society. In the Elizabethan and early Stuart period the theatre was the focal point of the age. Public life was shot through with theatricality—monarchs ruled with ostentatious pageantry, rank and status were defined in a rigid code of dress—while on the stages the tensions and contradictions working to change the nation were embodied and played out. More than any other form, the drama addressed itself to the total experience of its society. Playgoing was inexpensive, and the playhouse yards were thronged with apprentices, fishwives, labourers, and the like, but the same play that was performed to citizen spectators in the afternoon would often be restaged at court by night. The drama's power to activate complex, multiple perspectives on a single issue or event resides in its sensitivity to the competing prejudices and sympathies of this diversely minded audience.

Moreover, the theatre was fully responsive to the developing technical sophistication of nondramatic literature. In the hands of Shakespeare the blank verse employed for translation by the Earl of Surrey became a medium infinitely mobile between extremes of formality and intimacy, while prose encompassed both the control of Hooker and the immediacy of Nashe. This was above all a spoken drama, glorying in the theatrical energies of language. And the stage was able to attract the most technically accomplished writers of its day because it offered, uniquely, a literary career with some realistic prospect of financial return. The decisive event was the opening of the first purpose-built London playhouse in 1576, and during the next 70 years some 20 theatres more are known to have operated. The quantity and diversity of plays they commissioned is little short of astonishing.

Humanism and popular taste in the theatre

Theatres in London and the provinces. So the London theatres were a meeting ground of humanism and popular taste. They inherited, on the one hand, a tradition of humanistic drama current at court, the universities, and the Inns of Court (collegiate institutions responsible for legal education). This tradition involved the revival of classical plays and attempts to adapt Latin conventions to English, particularly to reproduce the type of tragedy, with its choruses, ghosts, and sententiously formal verse, associated with Seneca (10 tragedies by Seneca in English translation appeared in 1581). A fine example of the type is *Gorboduc* (1561), by Thomas Sackville and Thomas Norton, a tragedy based on British chronicle history that draws for Elizabeth's benefit a grave political moral about irresponsible government. It is also the first English play in blank verse. On the other hand, all the professional companies performing in London continued also to tour in the provinces, and the stage was never allowed to lose contact with its roots in country show, pastime, and festival. The simple moral scheme that pitted virtues against vices in the mid-Tudor interlude was never entirely submerged in more sophisticated drama, and the "Vice," the tricksy villain of the morality play, survives,

in infinitely more amusing and terrifying form, in Shakespeare's *Richard III*. Another survival was the clown or fool, apt at any moment to step beyond the play's illusion and share jokes directly with the spectators. The intermingling of traditions is clear in two farces, Nicholas Udall's *Ralph Roister Doister* (1553) and the anonymous *Gammer Gurton's Needle* (1559), in which academic pastiche is overlaid with country game; and what the popular tradition did for tragedy is indicated in Thomas Preston's *Cambises, King of Persia* (*c.* 1560), a blood and thunder tyrant play with plenty of energetic spectacle and comedy. A third tradition was that of revelry and masques, practiced at the princely courts across Europe and preserved in England in the witty and impudent productions of the schoolboy troupes of choristers who sometimes played in London alongside the professionals. An early play related to this kind is the first English prose comedy, Gascoigne's *Supposes* (1566), translated from a reveling play in Italian. Courtly revel reached its apogee in England in the ruinously expensive court masques staged for James I and Charles I, magnificent displays of song, dance, and changing scenery performed before a tiny aristocratic audience and glorifying the king. The principal masque writer was Ben Jonson, the scene designer Inigo Jones.

Professional playwrights. The first generation of professional playwrights in England was known collectively as the "university wits." Their nickname identifies their social pretensions, but their drama was primarily middle class, patriotic, and romantic. Their preferred subjects were historical or pseudo-historical, mixed with clowning, music, and love interest. At times plot virtually evaporated; George Peele's *Old Wives' Tale* (*c.* 1595) and Nashe's *Summer's Last Will and Testament* (1600) are simply popular shows, charming medleys of comic turns, spectacle, and song. Peele was a civic poet, and his serious plays are bold and pageant-like; *The Arraignment of Paris* (1584) is a pastoral entertainment, designed to compliment Elizabeth. Robert Greene's speciality was comical histories, interweaving a serious plot set among kings with comic action involving clowns. In his *Friar Bacon and Friar Bungay* (1594) and *James IV* (1598) the antics of vulgar characters complement but also criticize the follies of their betters. Only John Lyly, writing for the choristers, endeavoured to achieve a courtly refinement. His *Gallathea* (1584) and *Endimion* (1591) are fantastic comedies in which courtiers, nymphs, and goddesses make rarefied love in intricate, artificial patterns, the very stuff of courtly dreaming.

The university wits

Christopher Marlowe. Outshining all these is Christopher Marlowe, who alone realized the tragic potential inherent in the popular style, with its bombast and extravagance. His heroes are men of towering ambition who speak blank verse of unprecedented (and occasionally monotonous) elevation, their "high astounding terms" embodying the challenge that they pose to the orthodox norms and limitations of the societies they disrupt. In *Tamburlaine the Great* (two parts, published 1590) and *Edward II* (*c.* 1591; published 1594) traditional political orders are overwhelmed by conquerors and politicians who ignore the boasted legitimacy of weak kings; *The Jew of Malta* (*c.* 1589; published 1633) studies the man of business whose financial acumen and trickery give him unrestrained power; *The Tragical History of Dr. Faustus* (*c.* 1593; published 1604) shows the overthrow of a man whose learning and atheism threaten even God. The main focus of all these plays is on the uselessness of society's moral and religious sanctions against pragmatic, amoral will. They patently address themselves to the anxieties of an age being transformed by new forces in politics, commerce, and science; indeed, the sinister, ironic prologue to *The Jew of Malta* is spoken by Machiavelli. In his own time Marlowe was damned as atheist, homosexual, and libertine, and his plays remain disturbing because his verse makes theatrical presence into the expression of power, enlisting the spectators' sympathies on the side of his gigantic villain-heroes. His plays thus present the spectator with dilemmas that can be neither resolved nor ignored, and they articulate exactly the divided consciousness of their time. There is a similar effect in *The Spanish Tragedy* (*c.* 1591), by Marlowe's friend Thomas Kyd, an early "revenge tragedy" in which the hero seeks justice for the loss of his son but, in an unjust world, can achieve it only by taking the law into his own hands. Kyd's use of Senecan conventions (notably a ghost impatient for revenge) in a Christian setting expresses a genuine conflict of values, making the hero's success at once triumphant and horrifying.

Shakespeare's works. Above all other dramatists stands William Shakespeare, a supreme genius whom it is impossible to characterize briefly. Shakespeare is unequaled as poet and intellect, but he remains elusive. His capacity for assimilation—what Keats called his "negative capability"—means that his work is comprehensively accommodating; every attitude or ideology finds its resemblance there, yet also finds itself subject to criticism and interrogation. In part, Shakespeare achieved this by the total inclusiveness of his aesthetic, by putting clowns in his tragedies and kings in his comedies, juxtaposing public and private, and mingling the artful with the spontaneous; his plays imitate the counterchange of values occurring at large in his society. The sureness and profound popularity of his taste enabled him to lead the English Renaissance without privileging or prejudicing any one of its divergent aspects, while as actor, dramatist, and shareholder in the Lord Chamberlain's players he was involved in the Elizabethan theatre at every level. His career (dated from 1589 to 1613) was exactly coterminous with the period of greatest literary flourishing, and only in his work are the total possibilities of the Renaissance fully realized.

The early histories. Shakespeare's early plays were principally histories and comedies. About a fifth of all Elizabethan plays were histories, but this was the genre that Shakespeare particularly made his own, dramatizing the whole sweep of English history from Richard II to Henry VII in two four-play sequences, an astonishing project carried off with triumphant success. The first sequence, comprising the three *Henry VI* plays and *Richard III* (1589–92), begins as a patriotic celebration of English valour against the French. But this is soon superseded by a mature, disillusioned understanding of the world of politics, culminating in the devastating portrayal of Richard III—probably the first "character," in the modern sense, on the English stage—who boasts in *Henry VI,* Part 3, that he can "set the murtherous Machevil to school." Ostensibly *Richard III* monumentalizes the glorious accession of the dynasty of Tudor, but its realistic depiction of the workings of state power insidiously undercuts such platitudes, and the appeal of Richard's quick-witted individuality is deeply unsettling, short-circuiting any easy moral judgments. The second sequence, *Richard II* (1595), *Henry IV* (two parts, 1596–98), and *Henry V* (1599), begins with the deposing of a bad but legitimate king and follows its consequences through two generations, probing relentlessly at the difficult questions of authority, obedience, and order that it raises. (The Earl of Essex' faction paid for a performance of *Richard II* on the eve of their ill-fated rebellion against Elizabeth.) In the *Henry IV* plays, which are dominated by the massive character of Falstaff and his roguish exploits in Eastcheap, Shakespeare intercuts scenes among the rulers with scenes among those who are ruled to create a multifaceted composite picture of national life at a particular historical moment. The tone of these plays, though, is increasingly pessimistic, and in *Henry V* a patriotic fantasy of English greatness is hedged around with hesitations and qualifications about the validity of the myth of glorious nationhood offered by the Agincourt story. Through all these plays runs a concern for the individual and his subjection to historical and political necessity, a concern that is essentially tragic and anticipates greater plays yet to come. Shakespeare's other history plays, *King John* (*c.* 1591) and *Henry VIII* (1613) approach similar questions through material drawn from John Foxe's *Actes and Monuments.*

History sequences

The early comedies. The early comedies share the popular and romantic forms used by the university wits but overlay them with elements of elegant courtly revel and a sophisticated consciousness of comedy's fragility and artifice. These are festive comedies, giving access to a

Festivity, sportiveness, and the role of nature

society vigorously and imaginatively at play. One group, *The Comedy of Errors* (*c.* 1589–94), *The Taming of the Shrew* (*c.* 1590–94), *The Merry Wives of Windsor* (*c.* 1597–1601), and *Twelfth Night* (1601), are comedies of intrigue, fast moving, often farcical, and placing a high premium on wit. A second group, *The Two Gentlemen of Verona* (*c.* 1592–93), *Love's Labour's Lost* (*c.* 1595), *A Midsummer Night's Dream* (*c.* 1595–96), and *As You Like It* (1599), have as a common denominator a journey to a natural environment, such as a wood or park, in which the restraints governing everyday life are released and the characters are free to remake themselves untrammeled by society's forms, sportiveness providing a space in which the fragmented individual may recover wholeness. All the comedies share a belief in the positive, health-giving powers of play, but none is completely innocent of doubts about the limits that encroach upon the comic space, and in the four plays that approach tragicomedy, *The Merchant of Venice* (*c.* 1596–97), *Much Ado About Nothing* (1598–99), *All's Well That Ends Well* (1602–03), and *Measure for Measure* (1604), festivity is in direct collision with the constraints of normality, with time, business, law, human indifference, treachery, and selfishness. These plays give greater weight to the less optimistic perspectives on society current in the 1590s, and their comic resolutions are openly acknowledged to be only provisional, brought about by manipulation, compromise, or the exclusion of one or more major characters. The unique play *Troilus and Cressida* (*c.* 1601–03) presents a kind of theatrical no-man's-land between comedy and tragedy, between satire and savage farce. Shakespeare's reworking of the Trojan War pits heroism against its parody in a way that voices fully the fin-de-siècle sense of man's confused and divided individuality.

The tragedies. The confusions and contradictions of Shakespeare's age find their highest expression in his tragedies. In these extraordinary achievements, all values, hierarchies, and forms are tested and found wanting, and all society's latent conflicts are activated. Shakespeare sets husband against wife, father against child, the individual against society; he uncrowns kings, levels the nobleman with the beggar, and interrogates the gods. Already in the early experimental tragedies *Titus Andronicus* (*c.* 1592–94), with its spectacular violence, and *Romeo and Juliet* (*c.* 1595), with its comedy and romantic tale of adolescent love, Shakespeare had broken away from the conventional Elizabethan understanding of tragedy as a twist of fortune to an infinitely more complex investigation of character and motive, and in *Julius Caesar* (1599) he begins to turn the political interests of the history plays into secular and corporate tragedy, as men fall victim to the unstoppable train of public events set in motion by their private misjudgments. In the major tragedies that follow, Shakespeare's practice cannot be confined to a single general statement that covers all cases, for each tragedy belongs to a separate category: revenge tragedy in *Hamlet* (1600), domestic tragedy in *Othello* (*c.* 1603–04), social tragedy in *King Lear* (1605), political tragedy in *Macbeth* (1606), and heroic tragedy in *Antony and Cleopatra* (1607). In each category Shakespeare's play is exemplary and defines its type; the range and brilliance of this achievement is staggering. The worlds of Shakespeare's heroes are collapsing around them, and their desperate attempts to cope with the collapse uncover the inadequacy of the systems by which they rationalize and justify their existence. The ultimate insight is Lear's irremediable grief over his dead daughter: "Why should a dog, a horse, a rat, have life,/And thou no breath at all?" Before the overwhelming suffering of these great and noble spirits, all consolations are void and all versions of order stand revealed as adventitious. The humanism of the Renaissance is punctured in the very moment of its greatest single product.

Investigation of character and motive

Shakespeare's later works. In his last period, Shakespeare's astonishingly fertile invention returned to experimentation. In *Coriolanus* (1608) he completed his political tragedies, drawing a dispassionate analysis of the dynamics of the secular state; in the scene of the Roman food riot (not unsympathetically depicted) that opens the play is echoed the Warwickshire enclosure riots of 1607. *Timon of Athens* (1607–08) is an unfinished spin-off, a kind of tragical satire. The last group of plays comprises the four romances, *Pericles* (*c.* 1607–08), *Cymbeline* (*c.* 1609–10), *The Winter's Tale* (*c.* 1610–11), and *The Tempest* (1611), which develop a long, philosophical perspective on fortune and suffering. (A final work, *The Two Noble Kinsmen,* 1613, was written in collaboration with John Fletcher.) In these plays Shakespeare's imagination returns to the popular romances of his youth and dwells on mythical themes—wanderings, shipwrecks, the reunion of sundered families, and the resurrection of people long thought dead. There is consolation here, of a sort, beautiful and poetic, but still the romances do not turn aside from the actuality of suffering, chance loss, and unkindness, and Shakespeare's subsidiary theme is a sustained examination of the nature of his own art, which alone makes these consolations possible. Even in this unearthly context a subtle interchange is maintained between the artist's delight in his illusion and his mature awareness of his own disillusionment.

Playwrights after Shakespeare. Shakespeare's perception of a crisis in public norms and private belief became the overriding concern of the drama until the closing of the theatres in 1642. The prevailing manner of the playwrights who succeeded him was realistic, satirical, and antiromantic, and their comedies and tragedies focused predominantly on those two symbolic locations, the city and the court, with their typical activities, the pursuit of wealth and power. "Riches and glory," wrote Sir Walter Raleigh, "Machiavel's two marks to shoot at," had become the universal aims, and this situation was addressed by both "city comedy" and "tragedy of state." Increasingly, it was on the stages that the rethinking of early Stuart assumptions took place.

Theme of the pursuit of wealth and power

On the one hand, in the works of Thomas Heywood, Thomas Dekker, John Day, Samuel Rowley, and others, the old tradition of festive comedy was reoriented toward the celebration of confidence in the dynamically expanding commercial metropolis. Heywood claimed to have been involved in some 200 plays, and they include fantastic adventures starring citizen heroes, spirited, patriotic, and inclined to a leveling attitude in social matters. His masterpiece, *A Woman Kilde with Kindnesse* (1603), is a middle-class tragedy. Dekker was a kindred spirit, best seen in his *Shoemakers' Holiday* (1599), a celebration of citizen brotherliness and Dick Whittington-like success, which nevertheless faces squarely up to the hardships of work, thrift, and the contempt of the great. On the other hand, the very industriousness that the likes of Heywood viewed with civic pride became in the hands of Ben Jonson, George Chapman, John Marston, and Thomas Middleton a sign of aggression, avarice, and anarchy, symptomatic of the sicknesses in society at large.

Ben Jonson. The crucial innovations in satiric comedy were made by Ben Jonson, Shakespeare's friend and nearest rival, who stands at the fountainhead of what has subsequently been the dominant modern comic tradition. His early plays, particularly *Every Man in His Humour* (1598) and *Every Man Out of His Humour* (1599), with their galleries of grotesques, scornful detachment, and rather academic effect, were patently indebted to the verse satires of the 1590s; they introduced to the English stage a vigorous and direct anatomizing of "the time's deformities," the language, habits, and humours of the London scene. Jonson began as a self-appointed social legislator, aristocratic, conservative, and authoritarian, outraged by a society given over to inordinate appetite and egotism and ambitious through his mammoth learning to establish himself as the privileged artist, the fearless and faithful mentor and companion to kings; but he was ill at ease with a court inclined in its masques to prefer flattery to judicious advice. Consequently the greater satires that followed are marked by their gradual accommodations with popular comedy and by their unwillingness to make their implied moral judgments explicit: in *Volpone* (1606) the theatrical brilliance of the villain easily eclipses the sordid legacy hunters whom he deceives; *Epicoene* (1609) is a noisy farce of metropolitan fashion and frivolity; *The Alchemist* (1610) exhibits the conjurings and deceptions

The London scene in Jonson's plays

of clever London rogues; and *Bartholomew Fair* (1614) draws a rich portrait of city life parading through the annual fair at Smithfield, a vast panorama of a complete society. In these plays, fools and rogues are indulged to the very height of their daring, forcing upon the audience both criticism and admiration; the strategy leaves the audience to draw its own conclusions while liberating Jonson's wealth of exuberant comic invention, virtuoso skill with plot construction, and mastery of a language tumbling with detailed observation of London's multifarious ephemera. After 1616 Jonson abandoned the stage for the court, but, finding himself increasingly disregarded, he made a hard-won return to the theatres. The most notable of his late plays are popular in style: *The New Inn* (1629), which has affinities with the Shakespearean romance, and *A Tale of a Tub* (1633), which resurrects the Elizabethan country farce.

Marston and Middleton. Of Jonson's successors in city comedy, Francis Beaumont, in *The Knight of the Burning Pestle* (1607), amusingly insults the citizenry while ridiculing their taste for romantic plays. John Marston adopts so sharp a satirical tone that his plays in this genre frequently border on tragedy. All values are mocked by Marston's bitter and universal skepticism; his city comedy *The Dutch Courtezan* (1604), set in London, quotes a defense of libertinism from Montaigne. His tragicomedy *The Malcontent* (1604) is remarkable for its wild language and sexual and political disgust; Marston cuts the audience adrift from the moorings of reason by a dizzying interplay of parody and seriousness. Only in the city comedies of Thomas Middleton was Jonson's moral concern with greed and self-ignorance bypassed, for Middleton accepts the pursuit of money as, inevitably, the sole human absolute and presents buying and selling, usury, law, and the wooing of rich widows as the dominant modes of social interaction. His unprejudiced satire touches the actions of citizen and gentleman with equal irony and detachment; the only operative distinction is between fool and knave, and the sympathies of the audience are typically engaged on the side of wit, with the resourceful prodigal and dexterous whore. His characteristic form, used in *Michaelmas Terme* (1605) and *A Tricke to Catch the Old-One* (1606), was intrigue comedy, which enabled him to portray his society dynamically, as a mechanism in which each sex and class pursues its own selfish interests. He was thus concerned less to characterize the individual in depth than to examine the inequalities and injustices of the world that cause him to behave as he does. *The Roaring Girle* (*c.* 1608) and *A Chaste Maid in Cheapside* (1613) are the only Jacobean comedies to rival the comprehensiveness of *Bartholomew Fair,* but their social attitudes are opposed to Jonson's; the misbehaviour that Jonson condemned morally as "humours" or affectation Middleton understands as the product of circumstance.

Social concerns in Middleton's tragedies

Middleton's social concerns are also powerfully operative in his great tragedies, *Women Beware Women* (*c.* 1621) and *The Changeling* (1622), in which the moral complacency of men of rank is shattered by the dreadful violence they themselves have casually set in train, proving the answerability of all men for their actions despite the exemptions claimed for privilege and status. The hand of heaven is even more explicitly at work in the overthrow of the aristocratic libertine D'Amville in Cyril Tourneur's *Atheist's Tragedie* (*c.* 1611). Here the breakdown of old codes of deference before a progressive middle-class morality is strongly in evidence, and in *The Revenger's Tragedy* (1607), now generally attributed to Middleton, a scathing attack on courtly dissipation is reinforced by complaints about inflation and penury in the countryside at large. For more traditionally minded playwrights, new anxieties lay in the corrupt and sprawling bureaucracy of the modern court and in the political eclipse of the nobility before incipient royal absolutism. In Jonson's *Sejanus* (1603) Machiavellian statesmen abound, while George Chapman's *Bussy d'Ambois* (1604) and *Conspiracy of Charles, Duke of Byron* (1608) drew on recent French history to chart the collision of the magnificent but redundant heroism of the old-style aristocrat, whose code of honour had outlived its social function, with pragmatic arbitrary monarchy; Chapman doubtless had the career and fate of Essex in mind. The classic tragedies of state are John Webster's, with their dark Italian courts, intrigue and treachery, spies, malcontents, and informers. His *White Divel* (1612), a divided, ambivalent play, elicits sympathy even for a vicious heroine, since she is at the mercy of her deeply corrupt society; and the heroine in *The Duchess of Malfi* (1623) is the one decent and spirited inhabitant of her world, yet her noble death cannot avert the fearfully futile and haphazard carnage that ensues. As so often on the Jacobean stage, the challenge to the male-dominated world of power was mounted through the experience of its women.

Early Stuart drama. In the early Stuart period signs of a more polite drama, such as would prevail after 1660, were already beginning to appear in the comedies of fashionable manners written by John Fletcher and James Shirley, but even these playwrights lampooned courtiers and their overbearing ways. The traditions of a socially and politically critical theatre were carried down to the Civil War in the tragedies of John Ford (*'Tis Pitty Shee's a Whore,* 1633) and Philip Massinger (*Believe as You List,* 1631) and in comedies by Massinger (*A New Way to Pay Old Debts,* 1624; *The City Madam,* 1632) and Richard Brome (*The Antipodes,* 1638), which continued to probe at the tensions that were soon completely to undermine the basis of Stuart government. The outbreak of fighting in 1642 brought about the closing of the playhouses, but this was not because of any hostility by dramatists to politics or to change; rather, the crisis in which they were embroiled was one that had been the drama's continuing preoccupation for three generations.

EARLY STUART POETRY AND PROSE

In the early Stuart period the failure of consensus was dramatically announced in the political collapse of the 1640s and in the growing sociocultural divergences of the immediately preceding years. While it was still possible for the theatres to address the nation very much as a single audience, the court, with the baroque, absolutist style it encouraged in painting, masque, and panegyric, was becoming increasingly remote from the country at large and was regarded with justifiable distrust. In fact, a growing separation between polite and vulgar literature was to dispel many of the characteristic strengths of Elizabethan writing. Simultaneously, long-term intellectual changes were beginning to impinge on the status of poetry and prose. Sidney's defense of poetry, which maintained that poetry depicted what was ideally rather than actually true, was rendered redundant by the loss of agreement over transcendent absolutes; the scientist, the Puritan with his inner light, and the skeptic differed equally over the criteria by which truth or meaning was to be established. From the circle of Lord Falkland at Great Tew, which included poets such as Edmund Waller, Thomas Carew, and Sidney Godolphin, William Chillingworth argued that it was unreasonable for any individual to force his opinions onto any other, while Thomas Hobbes reached the opposite conclusion (in his *Leviathan,* 1651), that all must be as the state pleases. In this context, the old idea of poetry as a persuader to virtue fell obsolete, and the century as a whole witnessed a massive transfer of energy into new literary forms, particularly into the rationally balanced couplet, the autobiography, and the novel. At the same time, these influences were neither uniform nor consistent; Hobbes might repudiate the use of metaphor as senseless and ambiguous, yet his own prose is frequently enlivened by half-submerged metaphors.

Emergence of new literary forms

The Metaphysical poets. Writers responded to these conditions in different ways, and in poetry three types of practice may broadly be distinguished, which have been coupled with the names of Spenser, Jonson, and Donne. John Donne heads the tradition that Samuel Johnson typified for all time as the Metaphysicals; what unites them as a group is less the violent yoking of unlike ideas to which Johnson objected than that they were all poets of personal and individual feeling, responding to their time's pressures privately or introspectively (this very privateness, of course, was new; the period in general experienced a

massive trend toward contemplative or devotional verse).

Donne. Donne has been taken to be the apex of the 16th-century tradition of plain poetry, and certainly the love lyrics of his that parade their cynicism, indifference, and libertinism pointedly invert and parody the conventions of Petrarchan lyric, though no less than the Petrarchans he courts admiration for his poetic virtuosity. A "great haunter of plays" in his youth, he is always dramatic; his verse cultivates "strong lines," dissonance, and colloquiality. Thomas Carew praised him for exiling from poetry the "train of gods and goddesses"; what fills it instead is a dazzling battery of language and argument drawn from science, law and trade, court and city. Donne is the first London poet: his early satires and elegies are packed with the busy metropolitan milieu, and the songs and sonnets, which include his best writing, with their kaleidoscope of contradictory attitudes, ironies, and contingencies, are authentic to the modern phenomenon of urban living. Donne treats experience as relative, a matter of individual point of view; the personality is multiple, quizzical, and inconsistent, eluding definition. His love poetry is that of the frustrated careerist. By inverting normal perspectives and making the mistress "all states, and all princes, I, nothing else is," he belittles the public world, defiantly asserting the superior validity of his private experience, and frequently he erodes the traditional dichotomy of body and soul, outrageously praising the mistress in language reserved for platonic or religious contexts. The defiance is complicated, however, by a recurrent conviction of personal unworthiness that culminates in the *Anniversaries* (1611–12), two long commemorative poems written on the death of a patron's daughter. These expand into the classic statement of Jacobean melancholy, an intense meditation on the vanity of the world and the collapse of traditional certainties. Donne would, reluctantly, find respectability in a church career, but even his religious poems are torn between the same tense self-assertion and self-abasement that mark his secular poetry.

Donne's influence. Donne's influence was vast; the taste for wit and conceits reemerged in dozens of minor lyricists, among them courtiers such as Aurelian Townshend, William Habington, and William Cartwright and religious poets such as Francis Quarles and Henry King. The only true Metaphysical, in the sense of a poet with genuinely philosophical pretensions, was Edward Herbert (Lord Herbert of Cherbury), important as an early proponent of religion formulated by the light of reason. Donne's most interesting imitators were the three major religious poets—George Herbert, with his practical piety and richly domestic world, who substituted for Donne's tortured selfhood a humane, meditative assurance; the Roman Catholic Richard Crashaw, whose hymns introduced the sensuous ecstasies and effusions of the continental baroque; and Henry Vaughan, with his hermetic naturalism and mystical raptures.

In the context of the Civil War, however, Vaughan's and Crashaw's introspection begins to look like retreat, and when the satires of John Cleveland and the lyrics of Abraham Cowley take the Donne manner to extremes of paradox and vehemence, it suggests a loss of control in the face of political and social traumas. The one poet for whom metaphysical wit became a strategy for enforcing accommodations between conflicting allegiances was Donne's

Marvell

outstanding heir, Andrew Marvell. Marvell's finest writing is taut, extraordinarily dense and precise, uniquely combining a cavalier lyric grace with puritanical economy of statement. It seems to have been done at the time of greatest strain, in about 1650–53, and under the patronage of Sir Thomas Fairfax, parliamentarian general but opponent of the King's execution, whose retirement from politics to his country estate Marvell accorded qualified praise in "Upon Appleton House." His lyrics are poems of the divided mind, sensitive to all the major conflicts of their society—body against soul, action against retirement, experience against innocence, Oliver Cromwell against the King—but Marvell sustains the conflict of irreconcilables through paradox and wit rather than attempting to decide or transcend it. In this situation, irresolution has become a strength; in a poem like "An Horatian Ode upon Cromwell's Return from Ireland," which weighs the claims of King Charles and Cromwell, the poet's reserve was the only effective way of confronting the unprecedented demise of traditional structures of politics and morality.

Jonson and the Cavalier poets. By contrast, the Jonsonian tradition was, broadly, that of social verse, written with a classical clarity and weight and deeply informed by ideals of civilized reasonableness, ceremonious respect, and inner self-sufficiency derived from Seneca; it is a poetry of publicly shared values and norms. Jonson's own verse was occasional; it addresses other individuals, distributes praise and blame, and promulgates sober and judicious ethical attitudes. His favoured forms were the ode, elegy, satire, epistle, and epigram, and they are always crafted as exactly articulated objects, achieving a classical symmetry and monumentality. For Jonson the plain style meant not colloquiality but labour, restraint, and control; a good poet had first to be a good man, and his verses lead his society toward an aristocratic ethic of gracious but responsible living. With the Cavalier poets who succeeded him, the element of urbanity and conviviality tended to loom larger; Robert Herrick was perhaps England's first poet to express impatience with the tediousness of country life. However, Herrick's "The Country Life" and "The Hock Cart" rival Jonson's "To Penshurst" as panegyrics to the Horatian ideal of the "good life," calm and retired, but Herrick's poems gain poignancy by their implied contrast with the disruptions of the Civil War. The courtiers Thomas Carew, Sir John Suckling, and Richard Lovelace developed a manner of ease and naturalness suitable to the world of gentlemanly pleasure in which they moved; Suckling's *A Session of the Poets* lists more than 20 wits then in town. The Cavalier poets were writing England's first vers de société, lyrics of compliments and casual liaisons, often cynical, occasionally obscene; this was a line to be picked up again after 1660, as was the heroic verse and attitudinizing drama of Jonson's successor as poet laureate, Sir William Davenant. A different contribution was the elegance and smoothness that came to be associated with Sir John Denham and Edmund Waller, whom Dryden named as the first exponents of "good writing." Waller's polite lyrics now seem rather insipid, but Denham's topographical poem "Cooper's Hill" (1641), a considerable work in its own right, is plainly an important precursor of the balanced Augustan couplet (as is the otherwise slight oeuvre of Lucius Cary, Viscount Falkland). The growth of Augustan gentility was further encouraged by work done on translations in mid-century, particularly by Sir Richard Fanshawe (*Il Pastor Fido,* 1647) and Thomas Stanley.

Continued influence of Spenser. Donne had shattered Spenser's leisurely ornamentation, and Jonson censured his archaic language, but the continuing regard for Spenser at this time was significant. Variants of the Spenserian stanza were used by the brothers Giles and Phineas Fletcher, the former in his long religious poem *Christs Victorie* (1610), which is also indebted to Josuah Sylvester's highly popular translations from the French Calvinist poet Guillaume du Bartas, the *Divine Weeks and Works* (1605). Similarly, Spenserian pastorals still flowed from the pens of William Browne (*Britannia's Pastorals,* 1613–16), George Wither (*The Shepherd's Hunting,* 1614), and Michael Drayton, who at the end of his life returned nostalgically to portraying an idealized Elizabethan golden age (*The Muses Elizium,* 1630). Nostalgia was a dangerous quality under the progressive and absolutist Stuarts; the taste for Spenser involved a respect for values—traditional, patriotic, and Protestant—that were popularly, if erroneously, linked with the Elizabethan past but thought to be disregarded by the new regime. These poets believed they had a spokesman at court in the heroic and promising Prince Henry, but his death in 1612 disappointed many expectations, intellectual, political, and religious, and this group in particular was forced further toward the puritanical position. Increasingly their pastorals and fervently Protestant poetry aligned them in opposition to a court of Cavalier wits and of suspiciously pro-Spanish and pro-Catholic sympathies in foreign affairs; so sharp became Wither's satires that he earned imprisonment and was lampooned by Ben Jonson in a court masque. The failure

Literary opposition at court

of the Stuarts to conciliate attitudes such as these was to be crucial to their inability to maintain the cohesion of the Elizabethan compromise in the next generation. The nearest affinities, both in style and substance, of John Milton's early poetry would be with the Spenserians; in *Areopagitica* (1644) Milton praised "our sage and serious poet Spenser" as "a better teacher than Scotus or Aquinas."

Effect of religion and science on early Stuart prose. Puritanism also had a powerful effect on early Stuart prose. The best-sellers of the period were godly manuals that ran to scores of editions, like Arthur Dent's *Plain Man's Pathway to Heaven* (25 editions by 1640) and Lewis Bayly's *Practice of Piety* (1611; some 50 editions followed), the two of which formed the meagre dowry of John Bunyan's first wife. Puritans preferred sermons in the plain style too, eschewing rhetoric for an austerely profitable treatment of doctrine, though equally some famous godly preachers, such as Henry Smith and Thomas Adams, believed it their duty to make the Word of God eloquent. The other shaping factor was the desire among scientists for a utilitarian prose that would accurately and concretely represent the relationship between words and things, without figurative luxuriance. This hope, repeatedly voiced in the 1640s and '50s, eventually bore fruit in the practice of the Royal Society (incorporated 1662), which decisively affected prose after the Restoration. Its impact on earlier prose, though, was limited; most early Stuart science was written in the baroque style.

Sir Francis Bacon

The impetus toward a scientific prose derived ultimately from Sir Francis Bacon, the towering intellect of the century, who charted a philosophical system well in advance of his generation and beyond his own powers to complete. In the *Advancement of Learning* (1605) and the *Novum Organum* (1620) Bacon visualized a great synthesis of knowledge, rationally and comprehensively ordered so that each discipline might benefit from the discoveries of the others. The two radical novelties of his scheme were his insight that there could be progress in learning, that the limits of knowledge were not fixed but could be pushed forward, and his inductive method, by which scientific principles were to be established by experimentation, beginning at particulars and working toward generalities, instead of working backward from preconceived systems. Bacon democratized knowledge at a stroke, removing the tyranny of authority and lifting scientific inquiry free of religion and ethics and into the domain of mechanically operating second causes (though he held that the perfection of the machine itself testified to God's glory). The implications for prose are contained in his statement in the *Advancement* that the preoccupation with words instead of matter was the first "distemper" of learning; his own prose, however, was far from plain. The level exposition of idea in the *Advancement* is underpinned by a tactful but firmly persuasive rhetoric; and the famous *Essayes* (1597; enlarged 1612, 1625) are shifting and elusive, teasing the reader toward unresolved contradictions and half-apprehended complications.

The *Essayes* are masterworks in the new Stuart genre of the prose of leisure, the reflectively aphoristic prose piece in imitation of the *Essais* of Montaigne. Lesser collections were published by Sir William Cornwallis (1600–01), Owen Felltham (1623), and Ben Jonson (his posthumous *Timber; or, Discoveries*). A related genre was the "character," a brief, witty description of a social or moral type, imitated from Theophrastus, and practiced first by Joseph Hall (*Characters of Vertues and Vices,* 1608) and later by Sir Thomas Overbury, John Webster, and Thomas Dekker. The best characters are John Earle's (*Microcosmographie,* 1628). Character-writing led naturally into the writing of biography; the chief practitioners of this genre were Thomas Fuller, who included brief sketches in *The Holy State* (1642; includes *The Profane State*), and Izaak Walton, the biographer of Donne, George Herbert, and Richard Hooker. Walton's hagiographies are entertaining, but he manipulated the facts shamelessly; his biographies seem lightweight when placed beside Fulke Greville's tragical and valedictory *Life of the Renowned Sir Philip Sidney* (*c.* 1610; published 1652). The major historical work of the period was Sir Walter Raleigh's unfinished *History of the World* (1614), with its rolling periods and sombre skepticism, written from the Tower during his disgrace. Raleigh's providential framework would recommend his *History* to Cromwell and Milton; King James found it "too saucy in censuring princes." Bacon's *History of the Raigne of King Henry the Seventh* (1622) belongs to a more secular, Machiavellian tradition, which valued history for its lessons in pragmatism.

Prose styles. The essayists and character writers initiated a reaction against the orotund flow of serious Elizabethan prose that has been variously described as metaphysical, anti-Ciceronian, or Senecan, but these terms are used vaguely to denote both the cultivation of a clipped, aphoristic prose style, curt to the point of obscurity, and a fashion for looseness, asymmetry, and open-endedness. The age's professional stylists were the preachers, and in the sermons of Lancelot Andrewes and John Donne the clipped style is used to crumble the preacher's exegesis into tiny, hopping fragments or to suggest a nervous, agitated restlessness. An extreme example of the loose style is Robert Burton's *Anatomy of Melancholy* (1621), a massive encyclopaedia of learning, pseudoscience, and anecdote strung around an investigation into human psychopathology. Burton's compendiousness, his fascination with excess, necessitated a style that was infinitely extensible; his successor was Sir Thomas Urquhart, whose translation of *Gargantua and Pantagruel* (1653) outdoes even Rabelais. In the *Religio Medici* (1635), *The Garden of Cyrus,* and *Hydriotaphia, Urne-buriall, or A discourse of the Sepulchrall Urnes Lately Found in Norfolk* (1658) of Sir Thomas Browne the loose style serves a mind delighting in paradox and unanswerable speculation, content with uncertainty because of its intuitive faith in ultimate assurance. Browne's majestic prose invests his confession of his belief and his antiquarian and scientific tracts alike with an almost Byzantine richness and melancholy.

These were all learned styles, Latinate and sophisticated, but the appearance in the 1620s of the first *corantos,* or courants (news books), generated by interest in the Thirty Years' War, heralded the great 17th-century shift from an elite to a mass readership, a change effected by the explosion of popular journalism that accompanied the political confusion of the 1640s. The search for new kinds of political order and authority generated an answering chaos of styles, as voices were heard that had hitherto been denied access to print. The radical ideas of educated political theorists like Thomas Hobbes and the republican James Harrington were advanced within the traditional decencies of polite (if ruthless) debate, but they spoke in competition with vulgar writers who deliberately breached the literary canons of good taste—Levellers, such as John Lilburne and Richard Overton, with their vigorously dramatic manner; Diggers, like Gerrard Winstanley with his call for a general *Law of Freedom* (1652); and Ranters, whose language and syntax were as disruptive as the libertinism they professed. The outstanding examples were Milton's tracts against the bishops (1641–42), which revealed an unexpected talent for scurrilous abuse and withering sarcasm. Milton's later pamphlets, on divorce, education, and free speech (*Areopagitica,* 1644) and in defense of tyrannicide (*The Tenure of Kings and Magistrates,* 1649), adopt a loosely Ciceronian sonorousness, but their language is plain and always intensely imaginative and absorbing.

Rise of popular journalism

Milton's view of the poet's role. Milton had a concept of the public role of the poet even more elevated, if possible, than Jonson's; he early declared his hope to do for his native tongue what "the greatest and choicest wits of Athens, Rome, or modern Italy" had done for theirs. But where Jonson's humanism had led him toward a classical absolutism, Milton's was crossed by a respect for the conscience acting in pursuance of those things that it, individually, knew were right; he wished to "contribute to the progress of real and substantial liberty; which is to be sought for not from without, but within." His early verse aligned him, poetically and politically, with the Spenserians: religious and pastoral odes; "Lycidas" (1637), a pastoral elegy that incidentally bewails the state of the church; and *Comus* (1634), a masque against "masquing," performed privately in the country and opposing a private

heroism in chastity and virtue to the courtly round of revelry and pleasure.

During the interregnum, between the execution of Charles I and the restoration of Charles II, Milton saw his role as the intellectual serving the state in a glorious cause; he devoted his energies to pamphleteering, and he became Oliver Cromwell's Latin secretary. But the republic of virtue failed to materialize; Milton's courageous voice was the last before the Restoration to propose *The Ready and Easy Way to Establish a Free Commonwealth* (1660), a desperate program for a permanent oligarchy of the puritan elect, intended to avert the return to royal slavery. His greatest achievements, *Paradise Lost, Paradise Regained,* and *Samson Agonistes,* did not appear until several years after the Restoration, but their roots are deep in the radical experience of the 1640s and '50s and in the ensuing transformations in politics and society. For Milton and his contemporaries, 1660 was a watershed that was to necessitate a complete rethinking of expectations and ideas and a corresponding reassessment of the literary language, traditions, and forms appropriate to the new age. (M.H.B.)

The Restoration

LITERARY REACTIONS TO THE POLITICAL CLIMATE

The restoration of Charles II in 1660 led many to a painful revaluation of the political hopes and millenarian expectations bred during two decades of civil war and republican government. With the return of an efficient censorship, ambitiously heterodox ideas in theology and politics that had found their way freely into print during the 1640s and '50s were once again denied publication. The experience of defeat needed time to be absorbed, and fresh strategies had to be devised to encounter the challenge of hostile times. Much caustic and libelous political satire was written during the reigns of Charles II and James II and (because printing was subject to repressive legal constrictions) circulated anonymously and widely in manuscript. Andrew Marvell, sitting as member of Parliament for Hull in three successive parliaments from 1659 to 1678, experimented energetically with this mode, and his *Last Instructions to a Painter* (written in 1667) achieves a control of a broad canvas and an alertness to apt detail and to the movement of masses of people that make it a significant forerunner of Alexander Pope's *Dunciad* (however divergent the two poets' political visions may be). Marvell also proved himself to be a dexterous, abrasive prose controversialist, comprehensively deriding the anti-Dissenter arguments of Samuel Parker (later bishop of Oxford) in *The Rehearsal Transpros'd* (1672, with a sequel in 1673) and providing so vivid an exposition of Whig suspicions of the restored monarchy's attraction to absolutism in *An Account of the Growth of Popery, and Arbitrary Government in England* (1677) that a reward of £100 was offered for revealing its author's identity.

Return of censorship

The defeated republicans. The greatest prose controversialist of the pre-1660 years, John Milton, did not return to that mode but, in his enforced retirement from the public scene, devoted himself to his great poems of religious struggle and conviction, *Paradise Lost* (1667) and *Paradise Regained* and *Samson Agonistes* (both 1671). Each, in its probing of the intricate ways in which God's design reveals itself in human history, can justly be read (in one of its dimensions) as a chastened but resolute response to the failure of a revolution in which Milton himself had placed great trust and hope.

Others of the defeated republicans set out to record their own or others' experiences in the service of what they called the "good old cause." Lucy Hutchinson, for example, composed, probably in the mid-1660s, her remarkable memoirs of the life of her husband, Colonel Hutchinson, the Parliamentarian commander of Nottingham during the Civil War. Edmund Ludlow, like Hutchinson one of the regicides, fled to Switzerland in 1660, where he compiled his own *Memoirs.* These were published only in 1698–99 after Ludlow's death, and the discovery in 1970 of part of Ludlow's own manuscript revealed that they were edited and rewritten by another hand before printing. Civil War testimony still had political applications in the last years of the century, but those who sponsored its publication judged that Ludlow's now old-fashioned, millenarian rhetoric should be suppressed in favour of a soberer commonwealthman's dialect. Some autobiographers themselves adjusted their testimony in the light of later developments. George Fox, the Quaker leader, for example, dictating his *Journal* to various amanuenses, dubiously claimed for himself an attachment to pacifist principles during the 1650s, whereas it was, in fact, only in 1661, in the aftermath of the revolution's defeat, that the peace principle became central to Quakerism. The *Journal* itself only reached print in 1694 (again, after its author's death) after revision by a group superintended by William Penn. Such caution suggests a lively awareness of the influence such a text could have in consolidating a sect's sense of its own identity and continuity.

Writings of the Nonconformists. John Bunyan's *Grace Abounding* (1666), written while he was imprisoned in Bedford jail for nonconformity with the Church of England, similarly relates the process of his own conversion for the encouragement of his local, dissenter congregation. It testifies graphically to the force, both terrifying and consolatory, with which the biblical word could work upon the consciousness of a scantily educated, but overwhelmingly responsive, 17th-century believer. The form of *Grace Abounding* has numerous precedents in spiritual autobiography of the period, but with *The Pilgrim's Progress* (the first part of which appeared in 1678) Bunyan found himself drawn into a much more novel experiment, developing an ambitious allegorical narrative when his intent had been to write a more conventionally ordered account of the processes of redemption. The resulting work (with its second part appearing in 1684) combines a careful exposition of the logical structure of the Calvinist scheme of salvation with a delicate responsiveness to the ways in which his experience of his own world (of the life of the road, of the arrogance of the rich, of the rhythms of contemporary speech) can be deployed to render with a new vividness the strenuous testing the Christian soul must undergo. His achievement owes scarcely anything to the literary culture of his time, but his masterpiece has gained for itself a readership greater than that achieved by any other English 17th-century work with the exception of the King James Bible. Two other of his works, though lesser in stature, are especially worth reading: *The Life and Death of Mr. Badman* (1680), which, with graphic local detail, remorselessly tracks the sinful temptations of everyday life, and *The Holy War* (1682), a grandiose attempt at religious mythmaking interlaced with contemporary political allusions.

John Bunyan

Richard Baxter, a Nonconformist cleric who, although enduring persecution after 1660, was by instinct and much of his practice a reconciler, published untiringly on religious issues. He wrote, soon after the death of his wife, the moving *Breviate* (1681), a striking combination of exemplary narrative and unaffectedly direct reporting of the nature of their domestic life. His finest work, however, is the *Reliquiae Baxterianae* (published, five years after his death, in 1696), an autobiography that is also an eloquent defense of the Puritan impulse in the 17th-century Christian tradition.

The voice of anti-Puritan reaction can be heard in Samuel Butler's extensive mock-heroic satire *Hudibras* (published in three installments between 1662 and 1678). This was a massively popular work, with an influence stretching well into the 18th century (when Samuel Johnson, for example, greatly admired it and William Hogarth illustrated some scenes from it). It reads partly as a consummately destructive act of revenge upon those who had usurped power in the previous two decades, but although it is easy to identify what *Hudibras* opposes, it is difficult to say what, if anything, it affirms. Although much admired by Royalist opinion, it shows no wish to celebrate the authority or person restored in 1660, and its brazenly undignified use of rhyming tetrameters mirrors, mocks, and lacerates rooted human follies far beyond the power of one political reversal to obliterate. A comparable sardonic disenchantment is apparent in Butler's shorter verse satires and in his incisive and densely argued collection of prose *Characters.*

Samuel Butler

Writings of the Royalists. Royalists also resorted to biography and autobiography to record their experiences of defeat and restoration. Three of the most intriguing are by women: Margaret, duchess of Newcastle's life of her husband (1667) and the memoirs of Ann, Lady Fanshawe, and of Anne, Lady Halkett (both written in the late 1670s but not published in a fairly complete form until, respectively, 1829 and 1875). But incomparably the richest account of those years is *The History of the Rebellion and Civil Wars in England* by Edward Hyde, earl of Clarendon. The work was begun in exile during the late 1640s and was revised and completed in renewed exile after Clarendon's fall from royal favour in 1667. Clarendon was a close adviser to two kings, and his intimacy with many of the key events is unrivaled. Though his narrative is inevitably partisan, the ambitious range of his analysis and his mastery of character portraiture make the *History* an extraordinary accomplishment. His autobiography, which he also wrote during his last exile, gravely chronicles the transformations of the gentry world between the 1630s and '60s.

Role of the Church of England

In 1660 feeling in the country ran strongly in favour of the Church of England, persecution having confirmed in many a deep affection for Anglican rites and ceremonies. The reestablished church, accepting for itself the role of staunch defender of kingly authority, tended to eschew the exploration of ambitious and controversial theological issues and devoted itself instead to expounding codes of sound moral conduct. It was an age of eminent preachers (including Robert South, Isaac Barrow, Edward Stillingfleet, and John Tillotson) and of keen interest in the art of preaching. In conscious reaction against the obscurantist dialects judged typical of the sects, a plain and direct style of sermon oratory was favoured. Thus, in his funeral sermon on Tillotson in 1694, Gilbert Burnet praised the Archbishop because he "said what was just necessary to give clear Ideas of things, and no more" and "laid aside all long and affected Periods."

MAJOR GENRES AND MAJOR AUTHORS OF THE PERIOD

A comparable preference for an unembellished and perspicuous use of language is apparent in much of the nontheological literature of the age. Thomas Sprat, in his propagandizing *History of the Royal Society of London* (1667), and with the needs of scientific discovery in mind, also advocated "a close, naked natural way of speaking, positive expressions, clear senses, a native easiness." Sprat's work and a series of books by Joseph Glanvill, beginning with *The Vanity of Dogmatizing* (1661), argued the case for an experimental approach to natural phenomena against both the old scholastic philosophy and general conservative prejudice. That a real struggle was involved can be seen from the invariably disparaging attitude of contemporary satires to the labours of the Royal Society's enthusiasts (see, for instance, Samuel Butler's "The Elephant in the Moon," probably written in 1670–71, and Thomas Shadwell's *The Virtuoso,* 1676)—a tradition to be sustained later by Swift and Pope. But evidence of substantial achievement for the new generation of explorers was being published throughout the period, in, for example, Robert Boyle's *Sceptical Chymist* (1661), Robert Hooke's *Micrographia* (1665), John Ray's *Historia Plantarum* (in three volumes, 1686–1704), and, above all, Isaac (later Sir Isaac) Newton's *Philosophiae Naturalis Principia Mathematica* (1687).

Chroniclers. The Restoration, in its turn, bred its own chroniclers. Anthony à Wood, the Oxford antiquarian, made in his *Athenae Oxonienses* (1691–92) the first serious attempt at an English biographical dictionary. His labours were aided by John Aubrey, whose own unsystematic but enticing manuscript notes on the famous have been published in modern times under the title *Brief Lives.* After 1688 secret histories of the reigns of Charles II and James II were popular, of which the outstanding instance, gossipy but often reliable, is the *Memoirs of the Count Grammont,* compiled in French by Anthony Hamilton and first translated into English in 1714. A soberer but still free-speaking two-volume *History of My Own Time* (published posthumously, 1724–34) was composed by the industrious Gilbert Burnet, bishop of Salisbury from 1689. In the last months of the life of the court poet John Wilmot, 2nd earl of Rochester, Burnet had been invited to attend him, and in *Some Passages of the Life and Death of John, Earl of Rochester* (1680) he offered a fascinating account of their conversations as the erstwhile rake edged toward a rapprochement with the faith he had spurned.

A sparer, more finely focused prose was written by George Savile, 1st marquess of Halifax, who, closely involved in the political fray for 35 years, but remaining distrustful of any simple party alignments, wrote toward the end of his life a series of thoughtful, wryly observant essays, including *The Character of a Trimmer* (circulated in manuscript in late 1684 or very early 1685), *A Letter to a Dissenter* (published clandestinely in 1687), and *A Character of King Charles the Second* (written after about 1688). He also composed for his own daughter *The Lady's New-Year's-Gift; or, Advice to a Daughter* (1688), in which he anatomizes, with a sombre but affectionate wit, the pitfalls awaiting a young gentlewoman in life, especially in marriage.

Diarists. Two great diarists are among the most significant witnesses to the development of the Restoration world. Both possessed formidably active and inquisitive intelligences. John Evelyn was a man of some moral rectitude and therefore often unenamoured of the conduct he observed in court circles; but his curiosity was insatiable, whether the topic in question happened to be Tudor architecture, contemporary horticulture, or the details of sermon rhetoric. Samuel Pepys, whose diary, unlike Evelyn's, covers only the first decade of the Restoration, was the more self-scrutinizing of the two, constantly mapping his own behaviour with an alert and quizzical eye. Though not without his own moral inhibitions and religious gravity, Pepys immersed himself more totally than Evelyn in the new world of the 1660s, and it is he who gives the more resonant and idiosyncratic images of the changing London of the time.

The court wits. Among the subjects for gossip in London the group known as the "court wits" held a special place. Their conduct of their lives provoked censure from many, but among them were poets of some distinction, who drew upon the example of gentlemen-authors of the preceding generation (especially Sir John Suckling, Abraham Cowley, and Edmund Waller, the last two of whom themselves survived into the Restoration and continued to write impressive verse). The court wits' best works are mostly light lyrics, for example, Sir Charles Sedley's "Not, Celia, that I juster am" or Charles Sackville, earl of Dorset's "Dorinda's sparkling wit, and eyes." One of their number, the previously mentioned John Wilmot, earl of Rochester, possessed, however, a wider range and richer talent. Though some of his surviving poetry is in the least ambitious sense occasional work, he also produced writing of great force and authority, including a group of lyrics (for example, "All my past life is mine no more" and "An age in her embraces past") that, in psychological grasp and limpid deftness of phrasing, are among the finest of the century. He also wrote the harsh and scornfully dismissive *Satire Against Reason and Mankind* (probably before 1676) and experimented ingeniously with various forms of verse satire on contemporary society. The most brilliant of these, *A Letter from Artemisia in the Town, to Chloë in the Country* (written about 1675), combines a shrewd ear for currently fashionable idioms with a Chinese box structure that masks the author's own thoughts. Rochester's determined use of strategies of indirection anticipates Swift's tactics as an ironist.

John Oldham, a young schoolmaster, received encouragement as a poet from Rochester. His career, like his patron's, was to be cut short by an early death (in 1683, at age 30); but of his promise there can be no doubt. His *Satires upon the Jesuits* (1679–81), written during the Popish Plot, makes too unrelenting use of a rancorous, hectoring tone, but his development of the possibilities (especially satiric) of the "imitation" form, already explored by Rochester in, for example, *An Allusion to Horace* (written 1675–76), earns him an honourable place in the history of a mode that Pope was to put to such dazzling use. His imitation of the ninth satire of Horace's first

book exemplifies the agility and tonal resource with which Oldham could adapt a classical original to, and bring its values to bear upon, Restoration experience.

A poet who found early popularity with Restoration readers is Charles Cotton, whose *Scarronides* (1664–65), travesties of books one and four of Virgil's *Aeneid,* set a fashion for poetical burlesque. He is valued today, however, for work that attracted less contemporary interest but was to be admired by Wordsworth, Coleridge, and Charles Lamb. The posthumous *Poems on Several Occasions* (1689) includes deft poetry of friendship and love written with the familiar, colloquial ease of the Cavalier tradition and carefully observed, idiosyncratically executed descriptions of nature. He also added a second part to his friend Izaak Walton's *Compleat Angler* in 1676. A writer whose finest work was unknown to his contemporaries, much of it having been published only during the 20th century, is the poet and mystic Thomas Traherne. Influenced by the Hermetic writings and the lengthy Platonic tradition, he wrote, with extreme transparency of style, out of a conviction of the original innocence and visionary illumination of infancy. His poetry, though uneven, contains some remarkable writing, but his richest achievements are perhaps to be found in the prose *Centuries of Meditations* (first published in 1908).

Dryden. A poetic accomplishment of quite another order is that of John Dryden. He was 29 years old when Charles II returned from exile, and little writing by him survives from before that date. But for the remaining 40 years of his life he was unwearyingly productive, responding to the challenges of an unstable world with great formal originality and a mastery of many poetic styles. He was profoundly a poet of the public domain, but the ways in which he addressed himself to the issues of the day varied greatly in the course of his career. Thus, his poem to celebrate the Restoration itself, *Astraea Redux* (1660), invokes Roman ideas of the return of a golden age under Augustus Caesar in order to encourage similar hopes for England's future; whereas in 1681 the Exclusion Crisis (the attempt to exclude Charles II's brother James, a Roman Catholic, from succeeding to the throne) drew from Dryden one of his masterpieces, *Absalom and Achitophel,* in which the Old Testament story of King David, through an ingenious mingling of heroic and satiric tones, is made to shadow and comment decisively upon the current political confrontation. Another of his finest inventions, *Mac Flecknoe* (written mid-1670s, published 1682), explores, through agile mock-heroic fantasy, the possibility of a world in which the profession of humane letters has been thoroughly debased through the unworthiness of its practitioners. The 1680s also saw the publication of two major religious poems: *Religio Laici or a Laymans Faith* (1682), in which he uses a plain style to handle calmly the basic issues of faith, and *The Hind and the Panther* (1687), in which an elaborate allegorical beast fable is deployed to trace the history of animosities between Anglicanism and Roman Catholicism. In the Revolution of 1688 Dryden stayed loyal to the Catholicism to which he had been converted a few years earlier and thus lost his public offices. Financial need spurred him into even more literary activity thereafter, and his last years produced immensely skilled translations of Juvenal, Persius, and Virgil and handsome versions of Boccaccio and Chaucer, as well as further fine original poetry.

Dryden's commentary on public life

Dryden was also, in Samuel Johnson's words, the father of English criticism. Throughout his career he wrote extensively on matters of critical precept and poetic practice. Such sustained effort for which there was no precedent presumed the possibility of an interested audience but also contributed substantially to the creation of one. His tone is consistently exploratory and undogmatic. He writes as a working author, with an eye to problems he has himself faced, and is skeptical of theoretical prescriptions that threaten to become straitjackets for the poet or the critic. His discussion of Ben Jonson's *Epicoene, or The Silent Woman* in *Of Dramatick Poesie, an Essay* (1668) is remarkable as the first extended analysis of an English play, and his *Discourse Concerning the Origin and Progress of Satire* (1693) and the preface to the *Fables Ancient and Modern* (1700) both contain detailed commentary of the highest order.

Dryden as critic

A contrary critical philosophy was espoused by Thomas Rymer, an adherent of the most rigid neoclassical notions of dramatic decorum, who surveyed the pre-1642 English drama in *Tragedies of the Last Age* (1678) and *A Short View of Tragedy* (1693) and found it wanting. His zealotry reads unattractively today, but Dryden was impressed by him, if disinclined to accept his judgments without protest. In due course the post-1660 playwrights were to find their own scourge in Jeremy Collier, whose *Short View of the Immorality and Profaneness of the English Stage* (1698) comprehensively indicted the Restoration stage tradition. The theoretical frame of Collier's tract is crude, but his strength lay in his dogged citation of evidence from published play texts, especially when the charge was blasphemy, a crime still liable to stiff penalties in the courts. Even so clever a man as William Congreve was left struggling when attempting to deny in print the freedoms he had allowed his wit.

Drama by Dryden and others. Characteristically, Dryden, as dramatist, experimented vigorously in all the popular stage modes of the day, exploring the possibilities of the rhymed heroic play in the 1660s and early 1670s and producing some distinguished tragic writing in *All for Love* (1677) and *Don Sebastian* (1689); but his greatest achievement, *Amphitryon* (1690), is a comedy. In this he was typical of his age. Though there were individual successes in tragedy (especially Thomas Otway's *Venice Preserved,* 1682, and Nathaniel Lee's *Lucius Junius Brutus,* 1680), the splendour of the Restoration theatre lies in its comic creativity. Several generations of dramatists contributed to that wealth. In the 1670s the most original work can be found in Sir George Etherege's *Man of Mode* (1676), William Wycherley's *Country-Wife* (1675) and *Plain-Dealer* (1676), and Aphra Behn's two-part *Rover* (1677, 1681). Commentary has often claimed to detect a disabling repetitiveness in even the best Restoration comic invention, but an attentive reading of *The Country-Wife* and *The Man of Mode* will reveal how firmly the two authors, close acquaintances, have devised dramatic worlds significantly dissimilar in atmosphere that set distinctive challenges for their players. The disturbed years of the Popish Plot produced comic writing of matching mood, especially in Otway's abrasive *Soldier's Fortune* (1680) and Lee's extraordinary variation on the Madame de La Fayette novella, *The Princess of Cleve* (1681–82). After the Revolution of 1688 a series of major comedies hinged on marital dissension and questions (not unrelated to contemporary political traumas) of contract, breach of promise, and the nature of authority. These include, in addition to *Amphitryon,* Thomas Southerne's *Wives Excuse* (1691), Sir John Vanbrugh's *Relapse* (1696) and *Provok'd Wife* (1697), and George Farquhar's *Beaux Stratagem* (1707). These years also saw the premieres of Congreve's four comedies and one tragedy, climaxing with his masterpiece, *The Way of the World* (1700), a brilliant combination of intricate plotting and incisively humane portraiture. The pressures brought upon society at home by continental wars against the French also began to make themselves felt, the key text here being Farquhar's *Recruiting Officer* (1706), in which the worlds of soldier and civilian are placed in suggestive proximity.

Creativity in comic Restoration theatre

After 1710 contemporary writing for the stage waned in vitality. The 18th century is a period of great acting and strong popular enthusiasm for the theatre, but only a few dramatists (Gay, Fielding, Goldsmith, and Sheridan) achieved writing of a quality to compete with their predecessors' best, and even a writer of Richard Brinsley Sheridan's undeniable resource produced in his best plays—*The Rivals* (1775), *The School for Scandal* (1777), and *The Critic* (1779)—work that seems more like a technically ingenious, but cautious, rearrangement of familiar materials than a truly innovative contribution to the corpus of English comic writing for the stage. A number of the Restoration masterpieces, however, continued to be performed well into the new century, and the influence of this comic tradition was also strongly apparent in satiric poetry and the novel in the decades that followed.

Locke. One other late 17th-century figure with a formidable influence in the 18th century demands consideration: the philosopher John Locke. His *Essay Concerning Human Understanding* (1690) rejects a belief in innate ideas and argues that the mind at birth is a tabula rasa. Experience of the world can only be accumulated through the senses, which are themselves prone to unreliability. The *Essay,* cautiously concerned to define the exact limits of what the mind can truly claim to know, threw exciting new light on the workings of human intelligence and stimulated further debate and exploration through the fertility of its suggestions—for example, about the way in which ideas come to be associated. Locke was equally influential on political thought. He came from Puritan stock and was closely linked during the Restoration with leading Whig figures, especially the most controversial of them all, the Earl of Shaftesbury. His *Two Treatises of Government* (published in 1690, but mainly written during the Exclusion Crisis 10 years earlier) asserts the right of resistance to unjust authority and, in the last resort, of revolution. To establish this he had to think radically about the origins of civil society, the mutual obligations of subjects and rulers, and the rights of property. The resulting work became the crucial reference point from which subsequent debate took its bearings.

The 18th century

PUBLICATION OF POLITICAL LITERATURE

The expiry of the Licensing Act in 1695 halted state censorship of the press. During the next 20 years there were to be 10 general elections. These two factors combined to produce an enormous growth in the publication of political literature. Senior politicians, especially Robert Harley, saw the potential importance of the pamphleteer in wooing the support of a wavering electorate, and numberless hack writers produced copy for the presses. Richer talents also played their part. Harley, for instance, instigated Daniel Defoe's industrious work on the *Review* (1704–13), which consisted, in essence, of a regular political essay defending, if often by indirection, current governmental policy. He also secured Jonathan Swift's polemical skills for contributions to *The Examiner* (1710–11). Swift's most ambitious intervention in the paper war, again overseen by Harley, was *The Conduct of the Allies* (1711), a devastatingly lucid argument against any further prolongation of the War of the Spanish Succession. Writers like Defoe and Swift did not confine themselves to straightforward discursive techniques in their pamphleteering but experimented deftly with mock forms and invented personae to carry the attack home. According to contemporary testimony, Defoe's *Shortest-Way with the Dissenters* (1702) so brilliantly sustained its impersonation of a High Church extremist, its alleged narrator, that it was at first mistaken for the real thing. This avalanche of political writing whetted the contemporary appetite for reading matter generally and, in the increasing sophistication of its ironic and fictional maneuvers, assisted in preparing the way for the astonishing growth in popularity of narrative fiction during the subsequent decades.

Contribution to political reviews

Political journalism. After Defoe's *Review* the great innovation in periodical journalism came with the achievements of Richard Steele and Joseph Addison in *The Tatler* (1709–11) and then *The Spectator* (1711–12). In a familiar, easily approachable style they tackled a great range of topics, from politics to fashion, from aesthetics to the development of commerce. They aligned themselves with those who wished to see a purification of manners after the laxity of the Restoration and wrote extensively, with descriptive and reformative intent, about social and family relations. Their political allegiances were Whig, and in their creation of Sir Roger de Coverley they painted a wry portrait of the landed Tory squire as likable, possessed of good qualities, but feckless and anachronistic. Contrariwise, they spoke admiringly of the positive and honourable virtues bred by a healthy, and expansionist, mercantile community. Addison, the more original of the two, was an adventurous literary critic who encouraged esteem for the ballad through his enthusiastic account of *Chevy-Chase,* wrote a thoughtful and probing examen of *Paradise Lost,* and hymned the pleasures of the imagination in a series of papers deeply influential on 18th-century thought. The success with which Addison and Steele established the periodical essay as a prestigious form can be judged by the fact that they were to have more than 300 imitators before the end of the century. The awareness of their society and curiosity about the way it was developing, which they encouraged in their eager and diverse readership, left its mark on much subsequent writing.

Major political writers. *Pope.* Alexander Pope contributed to *The Spectator* and moved for a time in Addisonian circles; but from about 1711 onward his more influential friendships were with Tory intellectuals. His early verse shows a dazzling precocity, his *Essay on Criticism* (1711) combining ambition of argument with great stylistic assurance and *Windsor-Forest* (1713) achieving an ingenious, late Stuart variation on the 17th-century mode of topographical poetry. The mock-heroic *Rape of the Lock* (final version published in 1714) is an astonishing feat, marrying a rich range of literary allusiveness and a delicately ironic commentary upon the contemporary social world with a potent sense of suppressed energies threatening to break through the civilized veneer. That he could also write successfully in a more plaintive mode is shown by "Eloisa to Abelard" (1717), which, modeled on Ovid's heroic epistles, enacts with moving force Eloisa's struggle to reconcile grace with nature, virtue with passion. But the prime focus of his labours between 1713 and 1720 was his energetically sustained and scrupulous translation of Homer's *Iliad* (to be followed by the *Odyssey* in the mid-1720s). From that decade onward his view of the transformations wrought in Robert Walpole's England by economic individualism and opportunism grew increasingly embittered and despairing. In this he was following a common Tory trend, epitomized most trenchantly by the writings of his friend, the politician Henry St. John, 1st Viscount Bolingbroke. Pope's *Essay on Man* (1733–34) was a grand systematic attempt to buttress the notion of a God-ordained, perfectly ordered, all-inclusive hierarchy of created things. But his most probing and startling writing of these years comes in the four *Moral Essays* (1731–35), the series of Horatian imitations, and the final four-book version of *The Dunciad* (1743), in which he turns to anatomize with outstanding imaginative resource the moral anarchy and perversion of once-hallowed ideals he sees as typical of the commercial society in which he must perforce live.

Pope's *Rape of the Lock*

Thomson, Prior, and Gay. James Thomson also sided with the opposition to Walpole, but his poetry sustained a much more optimistic vision. In *The Seasons* (first published as a complete entity in 1730 but then massively revised and expanded until 1746) Thomson meditated upon, and described with fascinated precision, the phenomena of nature. He brought to the task a vast array of erudition and a delighted absorption in the discoveries of post-civil war, especially Newtonian, science, from whose vocabulary he borrowed freely. The image he developed of man's relationship to, and cultivation of, nature provided a buoyant portrait of the achieved civilization and wealth that ultimately derive from them and that, in his judgment, contemporary England enjoyed. The diction of *The Seasons* has many Miltonian echoes. In *The Castle of Indolence* (1748) Thomson's model is Spenserian, and its wryly developed allegory lauds the virtues of industriousness and mercantile achievement.

A poet who, at his best, chose a less ambitious song to sing is Matthew Prior, a diplomat and politician of some distinction, who essayed graver themes in *Solomon on the Vanity of the World* (1718), a disquisition on the vanity of human knowledge, but who also wrote some of the most direct and coolly elegant love poetry of the period. Prior's principal competitor as a writer of light verse was John Gay, whose *Trivia: or, the Art of Walking the Streets of London* (1716) catalogues the dizzying diversity of urban life through a dexterous burlesque of Virgil's *Georgics.* His *Fables,* particularly those in the 1738 collection, contain sharp, subtle writing, and his work for the stage, especially in *The What D'Ye Call It* (1715), *Three Hours After Mar-*

riage (1717; written with John Arbuthnot and Pope), and *The Beggar's Opera* (1728), shows a sustained ability to breed original and vital effects from witty generic cross-fertilization.

Swift. Swift, who also wrote verse of high quality throughout his career, like Gay favoured octosyllabic couplets and a close mimicry of the movement of colloquial speech. His technical virtuosity allowed him to switch assuredly from poetry of great destructive force to the intricately textured humour of *Verses on the Death of Dr. Swift* (completed in 1732; published 1739) and to the delicate humanity of his poems to Stella. But his prime distinction is, of course, as the greatest prose satirist in the English language. His period as secretary to the distinguished man of letters, Sir William Temple, gave him the chance to extend and consolidate his reading, and his first major work, *A Tale of a Tub* (1704), deploys its author's learning to chart the anarchic lunacy of its supposed creator, a Grub Street hack, whose solipsistic "modern" consciousness possesses no respect for objectivity, coherence of argument, or inherited wisdom from Christian or classical tradition. Techniques of impersonation were central to Swift's art thereafter. The *Argument Against Abolishing Christianity* (1708), for instance, offers brilliant ironic annotations on the "Church in Danger" controversy through the carefully assumed voice of a "nominal" Christian. That similar techniques could be adapted to serve specific political goals is demonstrated by "The Drapier's Letters" (1724–25), part of a successful campaign to prevent the imposition of a new, and debased, coinage on Ireland. Swift had hoped for preferment in the English church, but his destiny lay in Ireland, and the ambivalent nature of his relationship to that country and its inhabitants provoked some of his most demanding and exhilarating writing—above all, *A Modest Proposal* (1729), in which the ironic use of an invented persona achieves perhaps its most extraordinary and mordant development. His most wide-ranging satiric work, however, is also his most famous, *Gulliver's Travels* (1726). Swift grouped himself with Pope and Gay in hostility to the Walpole regime and the Hanoverian court, and that preoccupation leaves its mark on this work. But *Gulliver's Travels* also hunts larger prey. At its heart is a radical critique of human nature in which subtle ironic techniques work to part the reader from any comfortable preconceptions and challenge him to rethink from first principles his notions of man.

Swift's major works

Shaftesbury and others. More consoling doctrine was available in the popular writings of Anthony Ashley Cooper, 3rd earl of Shaftesbury, which were gathered in his *Characteristicks of Men, Manners, Opinions, Times* (1711). Although Shaftesbury had been tutored by Locke, he dissented from the latter's rejection of innate ideas and posited that man is born with a moral sense that is closely associated with his sense of aesthetic form. The tone of Shaftesbury's essays is characteristically idealistic, benevolent, gently reasonable, and unmistakably aristocratic. His optimism was buffeted by Bernard de Mandeville, whose *Fable of the Bees* (1714–29), which includes "The Grumbling Hire" (1705), takes a closer look at early capitalist society than Shaftesbury was prepared to do. Mandeville stressed the indispensable role played by the ruthless pursuit of self-interest in securing society's prosperous functioning. He thus favoured an altogether harsher view of man's natural instincts than Shaftesbury did and used his formidable gifts as a controversialist to oppose the various contemporary hypocrisies, philosophical and theological, that sought to deny the truth as he saw it. He was, in his turn, the target of acerbic rebukes by, among others, William Law, John Dennis, and Francis Hutcheson. George Berkeley, who criticized both Mandeville and Shaftesbury, set himself against what he took to be the age's irreligious tendencies and the obscurantist defiance by some of his philosophical forbears of the truths of common sense. His *Treatise Concerning the Principles of Human Knowledge* (1710) and *Three Dialogues Between Hylas and Philonous* (1713) continued the 17th-century debates about the nature of human perception, to which Descartes and Locke had contributed. The extreme lucidity and elegance of his style contrast markedly with the more effortful, but intensely earnest, prose of Joseph Butler's *Analogy of Religion* (1736), which also seeks to confront contemporary skepticism and ponders scrupulously the bases of man's knowledge of his creator. In a series of works beginning with *A Treatise of Human Nature* (1739–40), David Hume identified himself as a key spokesman for ironic skepticism and probed uncompromisingly the human mind's propensity to work by sequences of association and juxtaposition rather than by reason. Edmund Burke's *Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful* (1757) merged psychological and aesthetic questioning by hypothesizing that the spectator's or reader's delight in the sublime depended upon a sensation of pleasurable pain. An equally bold assumption about human psychology—in this case, that man is an ambitious, socially oriented, product-valuing creature—lies at the heart of Adam Smith's masterpiece of laissez-faire economic theory, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776).

THE NOVEL

The major novelists. *Defoe.* Such ambitious debates on society and human nature ran parallel with the explorations of a literary form finding new popularity with a large audience, the novel. Defoe, for example, fascinated by any intellectual wrangling, was always willing (amid a career of unwearying activity) to publish his own views on the matter currently in question, be it economic, metaphysical, educational, or legal. His lasting distinction, though earned in other fields of writing than the disputative, is constantly underpinned by the generous range of his curiosity. Only someone of his catholic interests could have sustained, for instance, the superb *Tour Thro' the Whole Island of Great Britain* (1724–27), a vivid, county-by-county review and celebration of the state of the nation. He brought the same diversity of enthusiasms into play in writing his novels. The first of these, *Robinson Crusoe* (1719), an immediate success at home and on the Continent, is a unique fictional blending of the traditions of Puritan spiritual autobiography with an insistent scrutiny of the nature of man as social creature and an extraordinary ability to invent a sustaining modern myth. *A Journal of the Plague Year* (1722) displays enticing powers of self-projection into a situation of which Defoe can only have had experience through the narrations of others, and both *Moll Flanders* (1722) and *Roxana* (1724) lure the reader into puzzling relationships with narrators the degree of whose own self-awareness is repeatedly and provocatively placed in doubt.

Defoe's major works

Richardson. The enthusiasm prompted by Defoe's best novels demonstrated the growing readership for innovative prose narrative. Samuel Richardson, a prosperous London printer, was the next major author to respond to the challenge. His *Pamela: or, Virtue Rewarded* (1740, with a less happy sequel in 1741), using (like all Richardson's novels) the epistolary form, tells a story of an employer's attempted seduction of a young servant woman, her subsequent victimization, and her eventual reward in virtuous marriage with the penitent exploiter. Its moral tone is self-consciously rigorous and proved highly controversial. Its main strength lies in the resourceful, sometimes comically vivid imagining of the moment-by-moment fluctuations of the heroine's consciousness as she faces her ordeal. Pamela herself is the sole letter writer, and the technical limitations are strongly felt, though Richardson's ingenuity works hard to mitigate them. But Pamela's frank speaking about the abuses of masculine and gentry power sounds the skeptical note more radically developed in Richardson's masterpiece, *Clarissa: or, the History of a Young Lady* (1747–48), which has a just claim to being considered the most reverberant and moving tragic fiction in the English novel tradition. *Clarissa* uses multiple narrators and develops a profoundly suggestive interplay of opposed voices. At its centre is the taxing soul debate and eventually mortal combat between the aggressive, brilliantly improvisatorial libertine Lovelace and the beleaguered Clarissa, maltreated and abandoned by her family but abiding sternly loyal to her own inner sense of probity. The tragic consummation that grows from this

involves an astonishingly ruthless testing of the psychological natures of the two leading characters. After such intensities, Richardson's final novel, *The History of Sir Charles Grandison* (1753–54), is perhaps inevitably a less ambitious, cooler work, but its blending of serious moral discussion and a comic ending ensured it an influence on his successors, especially Jane Austen.

Fielding. Henry Fielding turned to novel writing after a successful period as a dramatist, during which his most popular work had been in burlesque forms. His entry into prose fiction was also in that mode. *An Apology for the Life of Mrs. Shamela Andrews* (1741), a travesty of Richardson's *Pamela,* transforms the latter's heroine into a predatory fortune hunter who cold-bloodedly lures her booby master into matrimony. Fielding continued his quarrel with Richardson in *The History of the Adventures of Joseph Andrews* (1742), which also uses *Pamela* as a starting point but which, developing a momentum of its own, soon outgrows any narrow parodic intent. His hostility to Richardson's sexual ethic notwithstanding, Fielding was happy to build, with a calm and smiling sophistication, on the growing respect for the novel to which his antagonist had so substantially contributed. In *Joseph Andrews* and *The History of Tom Jones, a Foundling* (1749) Fielding openly brought to bear upon his chosen form a battery of devices from more traditionally reputable modes (including epic poetry, painting, and the drama). This is accompanied by a flamboyant development of authorial presence. Fielding the narrator buttonholes the reader repeatedly, airs critical and ethical questions for the reader's delectation, and urbanely discusses the artifice upon which his fiction depends. In the deeply original *Tom Jones* especially, this assists in developing a distinctive atmosphere of self-confident magnanimity and candid optimism. His fiction, however, can also cope with a darker range of experience. *The Life of Mr. Jonathan Wild the Great* (1743), for instance, uses a mock-heroic idiom to explore a derisive parallel between the criminal underworld and England's political elite, and *Amelia* (1751) probes with sombre precision images of captivity and situations of taxing moral paradox.

Authorial presence in Fielding's novels

Smollett. Tobias Smollett had no desire to rival Fielding as a formal innovator, and his novels consequently tend to be rather ragged assemblings of disparate incidents. But, although uneven in performance, all of them include extended passages of real force and idiosyncracy. His freest writing is expended on grotesque portraiture in which the human is reduced to fiercely energetic automatism. Smollett can also be a stunning reporter of the contemporary scene, whether the subject be a naval battle or the gathering of the decrepit at a spa. His touch is least happy when, complying too facilely with the gathering cult of sensibility, he indulges in rote-learned displays of emotionalism and good-heartedness. His most sustainedly invigorating work can perhaps be found in *The Adventures of Roderick Random* (1748), *The Adventures of Peregrine Pickle* (1751), and (an altogether more interesting encounter with the dialects of sensibility) *The Expedition of Humphry Clinker* (1771).

Sterne. An experiment of a radical and seminal kind is Laurence Sterne's *Tristram Shandy* (1759–67), which, drawing on a tradition of learned wit from Erasmus and Rabelais to Burton and Swift, provides a brilliant comic critique of the progress of the English novel to date. The focus of attention is shifted from the fortunes of the hero himself to the nature of his family, environment, and heredity, and dealings within that family offer repeated images of human unrelatedness and disconnection. Tristram, the narrator, is isolated in his own privacy and doubts how much, if anything, he can know certainly even about himself. Sterne is explicit about the influence of Lockean psychology on his writing, and the book, fascinated with the fictive energies of the imagination, is filled with characters reinventing or mythologizing the conditions of their own lives. It also draws zestful stimulus from a concern with the limitations of language, both verbal and visual, and teases an intricate drama out of Tristram's imagining of, and playing to, the reader's likely responses. Sterne's *Sentimental Journey Through France and Italy* (1768) similarly defies conventional expectations of what a travel book might be. An apparently random collection of scattered experiences, it mingles affecting vignettes with episodes in a heartier, comic mode, but coherence of imagination is secured by the delicate insistence with which Sterne ponders how the impulses of sentimental and erotic feeling are psychologically interdependent.

Minor novelists. The work of these five giants was accompanied by interesting experiments from a number of lesser novelists. Sarah Fielding, for instance, Henry's sister, wrote penetratingly and gravely about friendship in *The Adventures of David Simple* (1744, with a sequel in 1753). Charlotte Lennox in *The Female Quixote* (1752) and Richard Graves in *The Spiritual Quixote* (1773) responded inventively to the influence of Cervantes, also discernible in the writing of Fielding, Smollett, and Sterne. John Cleland's *Memoirs of a Woman of Pleasure* (known as *Fanny Hill;* 1748–49) chose a more contentious path; in his charting of a young girl's sexual initiation, he experiments with minutely detailed ways of describing the physiology of intercourse. In emphatic contrast, Henry Mackenzie's *Man of Feeling* (1771) offers an extremist, and rarefied, version of the sentimental hero, while Horace Walpole's *Castle of Otranto* (1765) somewhat laboriously initiated the vogue for Gothic fiction. William Beckford's *Vathek* (1786), Ann Radcliffe's *Mysteries of Udolpho* (1794), and Matthew Lewis' *Monk* (1796) are among the more distinctive of its successors. But the most engaging and thoughtful minor novelist of the period is Fanny Burney, who was also an evocative and self-revelatory diarist and letter writer. Her *Evelina* (1778) and *Camilla* (1796) in particular handle with independence of invention and emotional insight the theme of a young woman negotiating her first encounters with a dangerous social world.

POETS AND POETRY AFTER POPE

Eighteenth-century poetry after Pope produced nothing that can compete with achievements on the scale of *Clarissa* and *Tristram Shandy;* but much that was vital was accomplished. William Collins' *Odes on Several Descriptive and Allegoric Subjects* (1747), for instance, displays great technical ingenuity and a resonant insistence on the imagination and the passions as poetry's true realm. The odes also mine vigorously the potentiality of personification as a medium for poetic expression. In his *Elegy Written in a Country Churchyard* (1751), Thomas Gray revisited the terrain of such recent poems as Thomas Parnell's *Night-Piece on Death* (1722) and Robert Blair's poem *The Grave* (1743) and discovered a tensely humane eloquence far beyond his predecessors' powers. In later odes, particularly *The Progress of Poesy* (1757), Gray successfully sought close imitation of the original Pindaric form, even emulating Greek rhythms in English, while developing ambitious ideas about cultural continuity and renewal. Gray's fascination with the potency of primitive art (as evidenced in another great ode, *The Bard,* 1757) is part of a larger movement of taste, of which the contemporary enthusiasm for James Macpherson's alleged translations of Ossian (1760–63) is a further indicator.

Experimentation in poetic expression

Another eclectically learned and energetically experimental poet is Christopher Smart, whose renown rests largely on two poems. *Jubilate Agno* (written during confinement in various asylums between 1758/59 and 1763 but not published until 1939) is composed in free verse and experiments with applying the antiphonal principles of Hebrew poetry to English. *A Song to David* (1763) is a rhapsodic hymn of praise, blending enormous linguistic vitality with elaborate structural patterning. Both contain encyclopaedic gatherings of recondite and occult lore, numerous passages of which modern scholarship has yet to explicate satisfactorily, but the poetry is continually energized by minute alterations of tone, startling conjunctions of material, and a unique alertness to the mystery of the commonplace. Smart was also a superb writer of hymns, a talent in which his major contemporary rival was William Cowper in his *Olney Hymns* (1779). Both are worthy successors to the richly inventive work of Isaac Watts in the first half of the century. Elsewhere, Cowper can write with buoyant humour and satiric relaxation, as when, for instance, he

wryly observes from the safety of rural seclusion the evils of town life. But some of his most characterful poetry emerges from a painfully intense experience of withdrawal and isolation. His rooted Calvinism caused him periods of acute despair when he could see no hope of admission to salvation, a mood chronicled with grim precision in his masterly short poem *The Castaway* (written 1799). His most extended achievement is *The Task* (1785), an extraordinary fusion of disparate interests, working calmly toward religious praise and pious acceptance.

Burns. The 1780s brought publishing success to Robert Burns for his *Poems, Chiefly in the Scottish Dialect* (1786). Drawing on the precedents of Allan Ramsay and Robert Fergusson, Burns demonstrated how Scottish idioms and ballad modes could lend a new vitality to the language of poetry. Although born a poor tenant farmer's son, Burns had made himself well versed in English literary traditions, and his innovations were fully premeditated. His range is wide, from uninhibitedly passionate love songs to sardonic satires on moral and religious hypocrisy, of which the monologue *Holy Willie's Prayer* (written 1785) is an outstanding example. His work bears the imprint of the revolutionary decades in which he wrote, and recurrent in much of it are a joyful hymning of freedom, both individual and national, and an instinctive belief in the possibility of a new social order.

Goldsmith. Two other major poets, both of whom also achieved distinction in an impressive array of nondramatic modes, demand attention: Oliver Goldsmith and Samuel Johnson. Goldsmith's contemporary fame as a poet rested chiefly on *The Traveller* (1764), *The Deserted Village* (1770), and the incomplete *Retaliation* (1774). The last, published 15 days after his own death, is a dazzling series of character portraits in the form of mock epitaphs on a group of his closest acquaintances. *The Traveller,* a philosophical comparison of the differing national cultures of western Europe and the degrees of happiness their citizens enjoy, is narrated by a restless wanderer whose heart yet yearns after his own native land, where his brother still dwells. In *The Deserted Village* the experience is one of enforced exile, as an idealized village community is ruthlessly broken up in the interests of landed power. A comparable story of a rural idyll destroyed (though, this time, narrative artifice allows its eventual restoration) is at the centre of his greatly popular but tonally elusive novel, *The Vicar of Wakefield* (1766). He was also a deft and energetic practitioner of the periodical essay, contributing to at least eight journals between 1759 and 1773. His *Citizen of the World,* originally published in *The Public Ledger* in 1760–61, uses the device of a Chinese traveler whose letters home comment tolerantly but shrewdly on his English experiences. He also produced two stage comedies, one of which, *She Stoops to Conquer* (1773), is one of the few incontrovertible masterpieces of the theatre after the death of Farquhar in 1707.

Johnson's poetry and prose. Goldsmith belonged to the circle of a writer of still ampler range and outstanding intellect, Samuel Johnson. Pope recognized Johnson's poetical promise in *London* (1738), an invigorating reworking of Juvenal's third satire as a castigation of the decadence of contemporary Britain. His finest poem, *The Vanity of Human Wishes* (1749), also takes its cue from Juvenal, this time his 10th satire. It is a tragic meditation on the pitiful spectacle of human unfulfillment, which yet ends with an urgent prayer of Christian hope. But, great poet though he was, the lion's share of his formidable energies was expended on prose. From his early years in London he lived by his pen and gave himself unstintingly to satisfy the booksellers' demands. Yet he managed to sustain a remarkable coherence of ethical ambition and personal presence throughout his voluminous labours. His twice-weekly essays for *The Rambler* (1750–52), for instance, consistently show his powers at their fullest stretch, handling an impressive array of literary and moral topics with a scrupulous intellectual gravity and attentiveness. Many of the preoccupations of *The Vanity of Human Wishes* and the *Rambler* essays reappear in *Rasselas* (1759), which catalogues with profound resource the vulnerability of human philosophies of life to humiliation at the hands of life itself. His forensic brilliance can be seen in his relentless review of Soame Jenyns' *Free Inquiry into the Nature and Origin of Evil* (1757), which caustically dissects the latter's complacent attitude to human suffering, and his analytic capacities are evidenced at their height in the successful completion of two major projects, his innovative *Dictionary of the English Language* (1755) and the great edition of Shakespeare's plays (1765). His last years produced much political writing (including the humanely resonant *Thoughts on the Late Transactions Respecting Falkland's Islands,* 1771); the socially and historically alert *Journey to the Western Islands of Scotland,* 1775; and the consummate *Lives of the Poets,* 1779–81. The latter was the climax of 40 years' writing of poetical biographies, including the multifaceted *Account of the Life of Mr. Richard Savage* (1744). These last lives, covering the period from Cowley to the generation of Gray, show Johnson's mastery of the biographer's art of selection and emphasis and (together with the preface and notes to his Shakespeare edition) contain the most provocative critical writing of the century. Although his allegiances lay with neoclassical assumptions about poetic form and language, his capacity for improvisatory responsiveness to practice that lay outside the prevailing decorums should not be underrated. His final faith, however, in his own creative practice as in his criticism, was that the greatest art eschews unnecessary particulars and aims toward carefully pondered and ambitious generalization. The same creed was eloquently expounded by another member of the Johnson circle, Sir Joshua Reynolds, in his 15 *Discourses* (delivered to the Royal Academy between 1769 and 1790, but first published collectively in 1797).

Johnson as essayist, editor, and political writer

The other prime source of Johnson's fame, his reputation as a conversationalist of epic genius, rests on the detailed testimony of contemporary memorialists including Fanny Burney, Hester Lynch Piozzi, and Sir John Hawkins. But the key text is James Boswell's magisterial *Life of Samuel Johnson* (1791). This combines in unique measure a deep respect for its subject's ethical probity and resourceful intellect with a far from inevitably complimentary eye for the telling details of his personal habits and deportment. Boswell manifests rich dramatic talent and a precise ear for conversational rhythms in his re-creation, and orchestration, of the debates that lie at the heart of this great biography. Another dimension of Boswell's literary talent came to light in the 1920s and '30s when two separate hoards of unpublished manuscripts were discovered. In these he is his own subject of study. The 18th century had not previously produced much autobiographical writing of the first rank, though the actor and playwright Colley Cibber's flamboyant *Apology for the Life of Mr. Colley Cibber* (1740) and William Cowper's sombre *Memoir* (written about 1766, first published in 1816) are two notable exceptions. But the drama of Boswell's self-observations has a richer texture than either of these. In the *London Journal* especially (covering 1762–63, first published in 1950), he records the processes of his dealings with others and of his own self-imaginings with a sometimes unnerving frankness and a tough willingness to ask difficult questions of himself.

Boswell's *Life of Samuel Johnson*

Boswell narrated his experiences at the same time as, or shortly after, they occurred. Edward Gibbon, on the other hand, taking full advantage of hindsight, left in manuscript at his death six autobiographical fragments, all having much ground in common, but each telling a subtly different version of his life. Though he was in many ways invincibly more reticent than Boswell, Gibbon's successive explorations of his own history yet form a movingly resolute effort to see the truth clearly. These writings were undertaken after the completion of the great work of his life, *The History of the Decline and Fall of the Roman Empire* (1776–88). He brought to the latter an untiring dedication in the gathering and assimilation of knowledge, an especial alertness to evidence of human fallibility and failure, and a powerful ordering intelligence supported by a delicate sense of aesthetic coherence. His central theme—that the destruction of the Roman Empire was the joint triumph of barbarism and Christianity—is sustained with formidable ironic resource. (M.Co.)

The Romantic period

THE NATURE OF ROMANTICISM

As a term to cover the most distinctive writers who flourished in the last years of the 18th century and the first decades of the 19th, "Romantic" is indispensable but also a little misleading: there was no self-styled "Romantic movement" at the time, and the great writers of the period did not call themselves Romantics.

Many of the age's foremost writers thought that something new was happening in the world's affairs, nevertheless. Blake's affirmation in 1793 that "A new Heaven is begun . . . " was matched a generation later by Shelley's "The world's great age begins anew." "These, these shall give the world/Another heart, and other pulses" wrote Keats, referring to Rousseau and Wordsworth. Fresh ideals came to the fore: in particular the ideal of freedom, long cherished in England, was being extended to every range of human endeavour. As that ideal swept through Europe, it became natural to believe that the age of tyrants might soon end.

Role of individual feeling and thought

The feature most likely to strike a reader turning to the poets of the time after reading their immediate predecessors is the new role of individual feeling and thought. Where the main trend of 18th-century poetics had been to praise the general, to see the poet as a spokesman of society, addressing a cultivated and homogeneous audience and having as his end the conveyance of "truth," the Romantics found the source of poetry in the particular, unique experience. Blake's marginal comment on Sir Joshua Reynolds' *Discourses* expresses the position with characteristic vehemence: "to generalise is to be an idiot; to particularise is the alone distinction of merit." The poet was seen as an individual distinguished from his fellows by the intensity of his perceptions, taking as his basic subject matter the workings of his own mind. The implied attitude to an audience varied accordingly: although Wordsworth maintained that a poet did not write "for Poets alone, but for Men," for Shelley the poet was "a nightingale who sits in darkness and sings to cheer its own solitude with sweet sounds," and Keats declared "I never wrote one single line of Poetry with the least Shadow of public thought." Poetry was regarded as conveying its own truth; sincerity was the criterion by which it was to be judged. Provided the feeling behind it was genuine, the resulting creation must be valuable.

The emphasis on feeling—seen perhaps at its finest in the poems of Burns—was in some ways a continuation of the earlier "cult of sensibility"; and it is worth remembering that Pope praised his father as having known no language but the language of the heart. But feeling had begun to receive particular emphasis and is found in most of the Romantic definitions of poetry. Wordsworth called it "the spontaneous overflow of powerful feeling," and in 1833 John Stuart Mill defined "natural poetry" as "Feeling itself, employing Thought only as the medium of its utterance." It followed that the best poetry was that in which the greatest intensity of feeling was expressed, and hence a new importance was attached to the lyric. The degree of intensity was affected by the extent to which the poet's imagination had been at work; as Coleridge saw it, the imagination was the supreme poetic quality, a quasi-divine creative force that made the poet a godlike being. Romantic theory thus differed from the neoclassic in the relative importance it allotted to the imagination: Samuel Johnson had seen the components of poetry as "invention, imagination and judgement" but William Blake wrote: "One Power alone makes a Poet: Imagination, the Divine Vision." The judgment, or conscious control, was felt to be secondary; the poets of this period accordingly placed great emphasis on the workings of the unconscious mind, on dreams and reveries, on the supernatural, and on the childlike or primitive view of the world, this last being regarded as valuable because its clarity and intensity had not been overlaid by the restrictions of civilized "reason." Rousseau's sentimental conception of the "noble savage" was often invoked, and often by those who were ignorant that the phrase is Dryden's or that the type was adumbrated in the "poor Indian" of Pope's *Essay on Man.* A further sign of the diminished stress placed on judgment is the Romantic attitude to form: if poetry must be spontaneous, sincere, intense, it should be fashioned primarily according to the dictates of the creative imagination. Wordsworth advised a young poet, "You feel strongly; trust to those feelings, and your poem will take its shape and proportions as a tree does from the vital principle that actuates it." This organic view of poetry is opposed to the classical theory of "genres," each with its own linguistic decorum; and it led to the feeling that poetic sublimity was unattainable except in short passages.

Importance of the imagination

Hand in hand with the new conception of poetry and the insistence on a new subject matter went a demand for new ways of writing. Wordsworth and his followers, particularly Keats, found the prevailing poetic diction of the later 18th century stale and stilted, or "gaudy and inane," and totally unsuited to the expression of their perceptions. It could not be, for them, the language of feeling, and Wordsworth accordingly sought to bring the language of poetry back to that of common speech. His theories of diction have been allowed to loom too large in critical discussion: his own best practice very often differs from his theory. Nevertheless, when Wordsworth published his preface to *Lyrical Ballads* in 1800, the time was ripe for a change: the flexible diction of earlier 18th-century poetry had hardened into a merely conventional language and, with the notable exceptions of Blake and Burns, little first-rate poetry had been produced (as distinct from published) in Britain since the 1740s.

POETRY

Blake, Wordsworth, and Coleridge. Useful as it is to trace the common elements in Romantic poetry, there was little conformity among the poets themselves. It is misleading to read the poetry of the first Romantics—William Blake, Samuel Taylor Coleridge, and William Wordsworth, for example—as if it had been written primarily to express their feelings. Their concern was rather to change the intellectual climate of the age. Blake had been dissatisfied since boyhood with the current state of poetry and the drabness of contemporary thought. His early development of a protective shield of mocking humour with which to face a world in which science had become trifling and art inconsequential is visible in the satirical *An Island in the Moon* (written *c.* 1784–85); he then took the bolder step of setting aside sophistication in the visionary *Songs of Innocence* (1789). His desire for renewal encouraged him to view the outbreak of the French Revolution as a momentous event. Tradition has it that he openly wore the revolutionary red cockade in the streets of London. In powerful works, such as *The Marriage of Heaven and Hell* (1790–93) and *Songs of Experience* (1794), he attacked the hypocrisies of the age and the impersonal cruelties resulting from the dominance of analytic reason in contemporary thought. As it became clear that the ideals of the Revolution were not likely to be realized in his time, he renewed his efforts to revise his contemporaries' view of the universe and to construct a new mythology centred not in the God of the Bible but in Urizen, a figure of reason and law who he believed to be the true deity worshiped by his contemporaries. The story of Urizen's rise to provide a fortification against the chaos created by loss of a true human spirit was set out first in "Prophetic Books" such as *The First Book of Urizen* (1794) and then, more ambitiously, in the unfinished manuscript *Vala,* or *The Four Zoas,* written from about 1796 to about 1807.

Dissatisfaction with the intellectual climate

Later Blake shifted his poetic aim once more. Instead of attempting a narrative epic on the model of *Paradise Lost* he produced the more loosely organized visionary narratives of *Milton* (1804–08) and *Jerusalem* (1804–20) where, still using mythological characters, he portrayed the imaginative artist as the hero of society and forgiveness as the greatest human virtue.

Wordsworth and Coleridge, meanwhile, were exploring the implications of the Revolution more intricately. Neither could easily forget the excitement of the period immediately following its outbreak. Wordsworth, who lived in France in 1791–92 and fathered an illegitimate child

there, was distressed when, soon after his return, Britain declared war on the republic, dividing his allegiance. While sharing the horror of his contemporaries at the massacres in Paris, he knew at first hand the idealism and generosity of spirit to be found among the revolutionaries. For the rest of his career he was to brood on the implications of those events, trying to develop a view of humanity that would be faithful to his twin sense of the pathos of individual human fates and of the unrealized potentialities in humanity as a whole. The first factor emerges in his early manuscript poems "The Ruined Cottage" and "The Pedlar" (both to form part of the later *Excursion*); the second was developed from 1797, when he and his sister, Dorothy, with whom he was living in the west of England, were in close contact with Coleridge. Stirred simultaneously by Dorothy's immediacy of feeling, manifested everywhere in her *Journals* (written 1798–1803, published 1897), and by Coleridge's imaginative and speculative genius, he produced the poems collected in *Lyrical Ballads* (1798). The volume began with Coleridge's "Rime of the Ancient Mariner," continued with poems displaying delight in the powers of nature and the humane instincts of ordinary people, and concluded with the meditative "Lines written a few miles above Tintern Abbey," an attempt to set out his mature faith in nature and humanity.

His investigation of the relationship between nature and the human mind continued in the long autobiographical poem addressed to Coleridge and later entitled *The Prelude* (1805; revised continuously and published posthumously, 1850). Here he traced the value for a poet of having been a child "fostered alike by beauty and by fear" (in true Gothic style) by an upbringing in sublime surroundings. The poem also makes much of the work of memory, a theme that reaches its most memorable expression in the "Ode: Intimations of Immortality from Recollections of Early Childhood." In poems such as "Michael" and "The Brothers," by contrast, written for the second volume of *Lyrical Ballads* (1800), Wordsworth dwelt on the pathos and potentialities of ordinary lives.

Coleridge's poetic development during these years paralleled Wordsworth's. Having briefly brought together images of nature and the mind in "The Eolian Harp" (1796), he had devoted himself to more public concerns in poems of political and social prophecy, such as "Religious Musings" and "The Destiny of Nations." Becoming disillusioned with contemporary politics, however, and encouraged by Wordsworth, he turned back to the relationship between nature and the human mind. Poems such as "This Lime-Tree Bower My Prison," "The Nightingale," and "Frost at Midnight" (now sometimes called the "conversation poems" but entitled more accurately by Coleridge himself "Meditative Poems in Blank Verse") combine sensitive descriptions of nature with subtlety of psychological comment. "Kubla Khan" (1797, published 1816), a poem that Coleridge said came to him in "a kind of Reverie," opened a new vein of exotic writing, which he exploited further in the supernaturalism of "The Ancient Mariner" and the unfinished "Christabel." After his visit to Germany in 1798–99, however, renewed attention to the links between the subtler forces in nature and the human psyche bore fruit in letters and notebooks; simultaneously, his poetic output became sporadic. "Dejection: An Ode" (1802), another meditative poem, which first took shape as a letter to Sara Hutchinson, Wordsworth's sister-in-law, memorably describes the suspension of his "shaping spirit of Imagination."

Nature and the human mind in Coleridge's poetry

The work of both poets was directed back to national affairs during these years by the rise of Napoleon. In 1802 Wordsworth dedicated a number of sonnets to the patriotic cause. The death in 1805 of his brother John, who was serving as a sea captain, was a grim reminder that while he had been living in retirement as a poet others had been willing to sacrifice themselves for the public good. From this time the theme of duty was to be prominent in his poetry. His political essay *Concerning the Relations of Great Britain, Spain and Portugal . . . as Affected by the Convention of Cintra* (1809) agreed with Coleridge's periodical *The Friend* (1809–10) in deploring the decline of principle among statesmen. When *The Excursion* appeared in 1814 (the time of Napoleon's first exile), Wordsworth announced the poem as the central section of a longer projected work, *The Recluse.* This work was to be "a philosophical Poem, containing views of Man, Nature, and Society," and Wordsworth hoped to complete it by adding "meditations in the Author's own Person." The plan was not fulfilled, however, and *The Excursion* was left to stand in its own right as a poem of consolation for those who had been disappointed by the failure of French revolutionary ideals.

Both Wordsworth and Coleridge benefited from the advent in 1811 of the Regency, which brought a renewed interest in the arts. Coleridge's lectures on Shakespeare and literature became fashionable, his plays were briefly produced, and he gained further celebrity from the publication in 1816 of a volume of poems called *Christabel, Kubla Khan, A Vision: The Pains of Sleep. Biographia Literaria* (1817), the account of his own development, combined philosophy and literary criticism in a new way; the account was lastingly influential for the insights it contained. Coleridge settled at Highgate in 1816, and he was sought there as "the most impressive talker of his age" (in the words of the essayist William Hazlitt). His later religious writings made a considerable impact on the Victorians.

Coleridge's *Biographia Literaria*

Other poets of the early Romantic period. Several of the lesser poets of this generation were more popular in their own time. The somewhat insipid *Fourteen Sonnets* (1789) of William Lisle Bowles were received with enthusiasm by Coleridge and Wordsworth. Thomas Campbell is now chiefly remembered for his patriotic lyrics such as "Ye Mariners of England" and "The Battle of Hohenlinden" (1807) and for the critical preface to his *Specimens of the British Poets* (1819); Samuel Rogers has survived for his brilliant table talk (published 1856, after his death, as *Recollections of the Table-Talk of Samuel Rogers*), rather than for his poetry. One of the most popular poets of the day was Thomas Moore, whose *Irish Melodies* began to appear in 1807. His highly coloured Oriental fantasy *Lalla Rookh* (1817) was also immensely popular.

Robert Southey was closely associated with Wordsworth and Coleridge and was looked upon as a prominent member, with them, of the "Lake School" of poetry. His grandiose epic poems, such as *Thalaba the Destroyer* (1801) and *The Curse of Kehama* (1810), were successful in their own time, but his fame is based on his prose work—the vigorous *Life of Nelson* (1813), the *History of the Peninsular War* (1823–32), and his classic formulation of the children's tale "The Three Bears."

George Crabbe wrote poetry of another kind: his sensibility, his values, much of his diction, and his heroic couplet verse form belong very firmly to the 18th century. He differs from the earlier Augustans, however, in his subject matter, concentrating on realistic, unsentimental accounts of the life of the poor and the middle classes. He shows considerable narrative gifts in his collections of verse tales (in which he anticipates many short-story techniques) and great powers of description. His main works, *The Village* (1783), *The Borough* (1810), *Tales in Verse* (1812), and *Tales of the Hall* (1819), gained him great popularity in the earlier 19th century; after a long period of neglect he is widely recognized once more as a major poet.

The later Romantics: Shelley, Keats, and Byron. The poets of the next generation shared their predecessors' passion for liberty (now set in a new perspective by the Napoleonic wars) and were in a position to learn from their experiments. Percy Bysshe Shelley in particular was deeply interested in politics, coming early under the spell of the anarchistic views of William Godwin, whose *Enquiry Concerning Political Justice* had appeared in 1793. Shelley's revolutionary ardour, coupled with a zeal for the liberation of mankind and a passion for poetry, caused him to claim in his critical essay *A Defence of Poetry* (1821, published 1840) that "the most unfailing herald, companion, and follower of the awakening of a great people to work a beneficial change in opinion or institution, is poetry," and that poets are "the unacknowledged legislators of the world." This fervour burns throughout the early *Queen Mab* (1813), the long *Laon and Cythna*

Shelley's political zeal and passion for poetry

(retitled *The Revolt of Islam,* 1818), and the lyrical drama *Prometheus Unbound* (1820). Shelley saw himself at once as poet and prophet, as the fine "Ode to the West Wind" (1819) makes clear. Despite his firm grasp of practical politics, however, it is a mistake to look for concreteness in his poetry, where his concern is with subtleties of perception and with the underlying forces of nature: his most characteristic image is of sky and weather, of lights and fires. His poetic stance invites the reader to respond with similar outgoing aspiration. It adheres to the Rousseauistic belief in an underlying spirit in individuals, one truer to human nature itself than the behaviour evinced and approved by society. In that sense his material is transcendental and cosmic and his expression thoroughly appropriate. Possessed of great technical brilliance, he is, at his best, a poet of excitement and power.

John Keats, by contrast, was a poet so richly sensuous that his early work, such as *Endymion* (1818)—"a trial of my Powers of Imagination" he called it—could produce an over-luxuriant, cloying effect. As the program set out in his early poem "Sleep and Poetry" shows, however, Keats was also determined to discipline himself: even before February 1820, when he first began to cough blood, he may have known that he had not long to live, and he devoted himself to the expression of his vision with feverish intensity. He experimented with many kinds of poem: "Isabella" (published 1820), an adaptation of a tale by Boccaccio, is a tour de force of craftsmanship in its attempt to reproduce a medieval atmosphere. His epic fragment *Hyperion* (begun in 1818 and abandoned, published 1820; later begun again and published as *The Fall of Hyperion,* 1856) has a new spareness of imagery, but Keats soon found the style too Miltonic and decided to give himself up to what he called "other sensations." Some of these "other sensations" are found in the poems of 1819, Keats's annus mirabilis: "The Eve of St. Agnes" and the great odes, "To a Nightingale," "On a Grecian Urn," and "To Autumn." These, with the *Hyperion* poems, represent the summit of Keats's achievement, showing what has been called "the disciplining of sensation into symbolic meaning," the complex themes being handled with a concrete richness of detail. Study of his poems is incomplete without a reading of his superb letters, which show the full range of the intelligence at work in his poetry.

Keats's achievement

George Gordon, Lord Byron, who differed from Shelley and Keats in themes and manner, was at one with them in reflecting their shift toward "Mediterranean" themes. Having thrown down the gauntlet in his early poem *English Bards and Scotch Reviewers* (1809), in which he directed particular scorn at poems and poets of sensibility and sympathy and declared his own allegiance to Milton, Dryden, and Pope, he developed a poetry of dash and flair, in many cases with a striking hero. His two longest poems, *Childe Harold's Pilgrimage* (1812–18) and *Don Juan* (1819–24), his masterpiece, provided alternative personae for himself, the one a bitter and melancholy exile among the historic sites of Europe, the other a picaresque adventurer enjoying a series of amorous adventures. The gloomy and misanthropic vein was further mined in dramatic poems such as *Manfred* (1817) and *Cain* (1821), which helped to secure his reputation in Europe, but he is now remembered best for witty, ironic, and less portentous writings, such as *Beppo* (1818), in which he first used the ottava rima form. The easy, nonchalant, biting style developed there became a formidable device in *Don Juan* and in his satire on Southey, *The Vision of Judgment* (1822).

Minor poets of the later period. Of the lesser poets of this generation the best is undoubtedly John Clare, a Northamptonshire man of humble background. His natural simplicity and lucidity of diction, his intent observation, his almost classical poise, and the unassuming dignity of his attitude to life make him one of the most quietly moving of English poets. Thomas Lovell Beddoes, whose violent imagery and obsession with death and the macabre recall the Jacobean dramatists, represents an imagination at the opposite pole; considerable metrical virtuosity is displayed in the songs and lyrical passages from his oversensational tragedy *Death's Jest-Book* (begun 1825; published posthumously, 1850). Another minor writer who found inspiration in the 17th century was George Darley, some of whose songs from *Nepenthe* (1835) keep their place in anthologies. The comic writer Thomas Hood once enjoyed a great vogue but is now little read, although such poems of social protest as *The Song of the Shirt* (1843) and "The Bridge of Sighs," and the graceful *Plea of the Midsummer Fairies* (1827), are by no means negligible.

THE NOVEL: AUSTEN, SCOTT, AND OTHERS

At the turn of the century the Gothic mode, with its alternations between evocation of terror and appeal to sensibility, reached a peak of popularity with novels such as Ann Radcliffe's *Mysteries of Udolpho* (1794) and *The Italian* (1797) and Matthew Gregory Lewis' sensational *The Monk* (1796). These writers dealt with the supernatural and with human psychology far less adequately than did the poets, however, and appear to modern readers all the more shallow when compared with the great novelist Jane Austen. Her *Northanger Abbey* (begun in 1797; published posthumously, 1817) satirizes the Gothic novel, among other things, with complex irony; *Sense and Sensibility* (begun 1797; published 1811) mocks the contemporary cult of sensibility, while also displaying sympathetic understanding of the genuine sensitivities to which it appealed; *Pride and Prejudice* (begun 1796; published 1813) shows how sanity and intelligence can break through the opacities of social custom. The limitation suggested by her narrow range of settings and characters is illusory; working within these chosen limits, she observed and described very closely the subtleties of personal relationships, while also appealing to a sense of principle which, like Wordsworth and Coleridge, she believed to be threatened in a fragmenting and increasingly cosmopolitan society. These qualities come to full fruition in *Mansfield Park* (1814), *Emma* (1815), and *Persuasion* (1817). A master of dialogue, she wrote with economy, hardly wasting a word.

The nature of society in Scott's works

The underlying debate concerning the nature of society is reflected also in the novels of Sir Walter Scott. After his earlier success as a poet in such narrative historical romances as *The Lay of the Last Minstrel* (1805), *Marmion* (1808), and *The Lady of the Lake* (1810), he turned to prose and wrote more than 20 novels, several of which concerned heroes who were growing up, as he and his contemporaries had done, in a time of revolutionary turmoil. In the best, such as *Waverley* (1814), *Old Mortality* (1816), and *The Heart of Midlothian* (1818), he reconstructs the recent past of his country, Scotland, from still surviving elements. His stress on the values of gallantry, fortitude, and human kindness, along with his picture of an older society in which all human beings have a recognized standing and dignity, appealed to an England in which class divisions were exacerbated by the new industrialism. His historical romances were to inspire many followers in the emerging new nations of Europe. Thomas Love Peacock's seven novels, by contrast, are conversation pieces in which many of the pretensions of the day are laid bare in the course of witty, animated, and genial talk. *Nightmare Abbey* (1818) explores the extravagances of contemporary intellectualism and poetry; the more serious side of his satire is shown in such passages as Mr. Cranium's lecture on phrenology in *Headlong Hall* (1816). The Gothic mode was developed interestingly by Mary Wollstonecraft Shelley (the daughter of William Godwin), whose *Frankenstein; or, The Modern Prometheus* (1818) explores the horrific possibilities of new scientific discoveries, and Charles Robert Maturin, whose *Melmoth the Wanderer* (1820) has, with all its absurdity, a striking intensity. Among lesser novelists may be mentioned Maria Edgeworth, whose realistic didactic novels of the Irish scene inspired Scott; Susan Edmonstone Ferrier, a Scot with her own vein of racy humour; John Galt, whose *Annals of the Parish* (1821) is a minor classic; and James Hogg, remembered for his remarkable *Private Memoirs and Confessions of a Justified Sinner* (1824), a powerful story of Calvinism and the supernatural.

MISCELLANEOUS PROSE

The Romantic emphasis on individualism is reflected in much of the prose of the period, particularly in criti-

cism and the familiar essay. Among the most vigorous, forthright, and least mannered writing is that of William Hazlitt, an energetic, enthusiastic, and subjective critic whose most characteristic work is seen in his collections of lectures *On the English Poets* (1818) and *On the English Comic Writers* (1819) and in *The Spirit of the Age* (1825), a series of valuable portraits of his contemporaries. In *The Essays of Elia* (1823) and *The Last Essays of Elia* (1833) Charles Lamb, an even more personal essayist, projects, with apparent artlessness, a carefully managed portrait of himself—charming, whimsical, witty, sentimental, warm-hearted, nostalgic, and sociable; as his fine *Letters* show, however, he could on occasion produce mordant satire. Thomas De Quincey also appealed to the new interest in personal writing, producing a colourful account of his early experiences in *Confessions of an English Opium Eater* (1821, revised and enlarged in 1856). His unusual gift of evoking states of dream and nightmare is best seen in essays such as "The English Mail Coach" and "On the Knocking at the Gate in *Macbeth*" and in his various autobiographical pieces.

Hazlitt, Lamb, and De Quincey

Of writers who might be called surviving classicists, the most notable is Walter Savage Landor, whose detached, lapidary style is seen at its best in some brief lyrics and in a series of erudite *Imaginary Conversations,* which began to appear in 1824. The anti-Romantic point of view received its most pungent expression in the pages of the journals: the Whig quarterly *Edinburgh Review* (begun 1802), edited by Francis Jeffrey, was followed by its Tory rivals *The Quarterly Review* (begun 1809) and the monthly *Blackwood's Magazine* (begun 1817). Their criticism was by no means always unjust and summed up much contemporary opinion; but the reviewers were too willing to judge the new poetry by their own settled standards, missing what was genuinely innovative. In their attacks on many kinds of prejudice and abuse, on the other hand, they set a notable standard of fearless and independent journalism. Similar independence was shown by Leigh Hunt, whose outspoken journalism, particularly in his *Examiner* (begun 1808), was of considerable influence, and by William Cobbett, whose *Rural Rides* (collected in 1830 from his *Political Register*) gives a telling picture, in forceful and clear prose, of the English countryside of his day.

DRAMA

Despite the unusually strong interest in the theatre, little drama of note emerged at this time. Most major poets produced plays, but although Coleridge's *Osorio* and *Zapolya* were produced in 1813 and 1818, respectively, and Byron's *Marino Falieri* in 1821, the achievements were literary rather than dramatic. At the Theatre Royal in Drury Lane where the acting of John Philip Kemble and his sister, Sarah Siddons, had been much admired, the centre of attention from 1814 onward was Edmund Kean, whose impassioned performances captivated Keats, Hazlitt, and Byron and of whom Coleridge said "To see him act is like reading Shakespeare by flashes of lightning." Coleridge's lectures and notes, which, along with the essays of Lamb and Hazlitt, brought a psychological and historical approach to Shakespeare and other early dramatists, set new standards of dramatic criticism during the period.

(R.P.C.M./J.B.B.)

The Post-Romantic and Victorian eras

Self-consciousness was the quality that John Stuart Mill identified, in 1838, as "the daemon of the men of genius of our time." Introspection was inevitable in the literature of an immediately Post-Romantic period, and the age itself was as prone to self-analysis as were its individual authors. William Hazlitt's essays *The Spirit of the Age* (1825) were echoed by Mill's articles of the same title in 1831, by Thomas Carlyle's essays "Signs of the Times" (1829) and "Characteristics" (1831), and by Richard Henry Horne's *New Spirit of the Age* in 1844.

Introspection in the literature of the age

This persistent scrutiny was the product of an acute sense of change. Britain had emerged from the long war with France (1793–1815) as a great power and as the world's predominant economy. Visiting England in 1847, the American writer Ralph Waldo Emerson observed of the English that "The modern world is theirs. They have made and make it day by day."

This new status as the world's first urban and industrialized society was responsible for the extraordinary wealth, vitality, and self-confidence of the period. Abroad these energies expressed themselves in the growth of the British Empire. At home they were accompanied by rapid social change and fierce intellectual controversy.

The juxtaposition of this new industrial wealth with a new kind of urban poverty is only one of the paradoxes that characterize this long and diverse period. In religion the climax of the Evangelical revival coincided with an unprecedentedly severe set of challenges to faith. In politics a widespread commitment to economic and personal freedom was, nonetheless, accompanied by a steady growth in the power of the state. The prudery for which the Victorian Age is notorious in fact went hand in hand with an equally violent immoralism, seen, for example, in Algernon Charles Swinburne's poetry or the writings of the Decadents. Most fundamentally of all, the rapid change that many writers interpreted as progress inspired in others a fierce nostalgia. Enthusiastic rediscoveries of ancient Greece, Elizabethan England, and, especially, the Middle Ages by writers, artists, architects, and designers made this age of change simultaneously an age of active and determined historicism.

John Stuart Mill caught this contradictory quality, with characteristic acuteness, in his essays on Jeremy Bentham (1838) and Samuel Taylor Coleridge (1840). Every contemporary thinker, he argued, was indebted to these two "seminal minds." Yet Bentham, as the enduring voice of the Enlightenment, and Coleridge, as the chief English example of the Romantic reaction against it, held diametrically opposed views.

A similar sense of sharp controversy is given by Carlyle in *Sartor Resartus* (1833–34). An eccentric philosophical fiction in the tradition of Swift and Sterne, the book argues for a new mode of spirituality in an age that Carlyle himself suggests to be one of mechanism. Carlyle's choice of the novel form and the book's humour, generic flexibility, and political engagement point forward to distinctive characteristics of Victorian literature.

EARLY VICTORIAN LITERATURE: THE AGE OF THE NOVEL

Several major figures of English Romanticism lived on into this period. Coleridge died in 1834, De Quincey in 1859. Wordsworth succeeded Southey as poet laureate in 1843 and held the post until his own death seven years later. Posthumous publication caused some striking chronological anomalies. Shelley's "Defence of Poetry" was not published until 1840. Keats's letters appeared in 1848 and Wordsworth's *Prelude* in 1850.

Despite this persistence critics of the 1830s felt that there had been a break in the English literary tradition, which they identified with the death of Byron in 1824. The deaths of Jane Austen in 1817 and Sir Walter Scott in 1832 should perhaps have been seen as even more significant, for the new literary era has, with justification, been seen as the age of the novel.

Dickens. Charles Dickens first attracted attention with the descriptive essays and tales originally written for newspapers, beginning in 1833, and collected as *Sketches by "Boz"* (1836). On the strength of this volume Dickens contracted to write a historical novel in the tradition of Scott (eventually published as *Barnaby Rudge* in 1841). By chance his gifts were turned into a more distinctive channel. In February 1836 he agreed to write the text for a series of comic engravings. The unexpected result was *The Pickwick Papers* (1836–37), one of the funniest novels in English literature. By July 1837 sales of the monthly installments exceeded 40,000 copies. Dickens' extraordinary popular appeal and the enormous imaginative potential of the Victorian novel were simultaneously established.

The chief technical features of Dickens' fiction were also formed by this success. Serial publication encouraged the use of multiple plot and required that each episode be individually shaped. At the same time it produced an unprecedentedly close relationship between author and

Serial publication of Dickens' work

reader. Part dramatist, part journalist, part mythmaker, and part wit, Dickens took the picaresque tradition of Smollett and Fielding and gave it a Shakespearean vigour and variety.

His early novels have been attacked at times for sentimentality, melodrama, or shapelessness. They are now increasingly appreciated for their comic or macabre zest and their poetic fertility. *Dombey and Son* (1846–48) marks the beginning of Dickens' later period. He thenceforth combined his gift for vivid caricature with a stronger sense of personality, designed his plots more carefully, and used symbolism to give his books greater thematic coherence. Of the masterpieces of the next decade, *David Copperfield* (1849–50) uses the form of a fictional autobiography to explore the great Romantic theme of the growth and comprehension of the self. *Bleak House* (1852–53) addresses itself to law and litigiousness, *Hard Times* (1854) is a Carlylian defense of art in an age of mechanism, and *Little Dorrit* (1855–57) dramatizes the idea of imprisonment, both literal and spiritual. Two great novels, both involved with issues of social class and human worth, appeared in the 1860s: *Great Expectations* (1860–61) and *Our Mutual Friend* (1864–65). His final book, *The Mystery of Edwin Drood* (published posthumously, 1870), was left tantalizingly uncompleted at the time of his death.

Thackeray, Gaskell, and others. Unlike Dickens, William Makepeace Thackeray came from a wealthy and educated background. The loss of his fortune at age 22, however, meant that he too learned his trade in the field of sketch writing and occasional journalism. His early fictions were published as serials in *Fraser's Magazine* or as contributions to the great Victorian comic magazine *Punch* (founded 1841). For his masterpiece, *Vanity Fair* (1847–48), however, he adopted Dickens' procedure of publication in monthly parts. Thackeray's satirical acerbity is here combined with a broad narrative sweep, a sophisticated self-consciousness about the conventions of fiction, and an ambitious historical survey of the transformation of English life in the years between the Regency and the mid-Victorian period. His later novels never match this sharpness. *Vanity Fair* was subtitled "A Novel Without a Hero." Subsequently, it has been suggested, a more sentimental Thackeray wrote novels without villains.

Elizabeth Gaskell began her career as one of the "Condition of England" novelists of the 1840s, responding like Frances Trollope, Benjamin Disraeli, and Charles Kingsley to the economic crisis of that troubled decade. *Mary Barton* (1848) and *Ruth* (1853) are both novels about social problems, as is *North and South* (1854–55), although, like her later work, this book also has a psychological complexity that anticipates George Eliot's novels of provincial life.

Variety of subgenres

Political novels, religious novels, historical novels, sporting novels, Irish novels, crime novels, and comic novels all flourished in this period. The years 1847–48, indeed, represent a pinnacle of simultaneous achievement in English fiction. In addition to *Vanity Fair, Dombey and Son,* and *Mary Barton,* they saw the completion of Disraeli's trilogy of political novels—*Coningsby* (1844), *Sybil* (1845), and *Tancred* (1847)—and the publication of first novels by Anne, Charlotte, and Emily Brontë; Charles Kingsley; and Anthony Trollope. For the first time literary genius appeared to be finding its most natural expression in prose fiction, rather than in poetry or drama. By 1853 the poet Arthur Hugh Clough would concede that "the modern novel is preferred to the modern poem."

The Brontës. In many ways, however, the qualities of Romantic verse could be absorbed, rather than simply superseded, by the Victorian novel. This is suggested clearly by the work of the Brontë sisters. Growing up in a remote but cultivated vicarage in Yorkshire, they invented, as children, the imaginary kingdoms of Angria and Gondal. These inventions supplied the context for many of the poems in their first, and pseudonymous, publication, *Poems by Currer, Ellis and Acton Bell* (1846). Their Gothic plots and Byronic passions also informed the novels that began to be published in the following year.

Charlotte Brontë, like her sisters, appears at first sight to have been writing a literal fiction of provincial life. In her first novel, *Jane Eyre* (1847), for example, the heroine's choice between sexual need and ethical duty belongs very firmly to the mode of moral realism. But her hair's-breadth escape from a bigamous marriage with her employer, and the death by fire of his mad first wife derive from the rather different tradition of the Gothic novel. In *Shirley* (1849) Charlotte Brontë strove to be, in her own words, "as unromantic as Monday morning." In *Villette* (1853) the distinctive Gothic elements return to lend this study of the limits of stoicism an unexpected psychological intensity and drama.

Emily Brontë united these diverse traditions still more successfully in her only novel, *Wuthering Heights* (1847). Closely observed regional detail, precisely handled plot, and a sophisticated use of multiple internal narrators are combined with vivid imagery and an extravagantly Gothic theme. The result is a perfectly achieved study of elemental passions and the strongest possible refutation of the assumption that the age of the novel must also be an age of realism.

EARLY VICTORIAN VERSE

Tennyson. Despite the growing prestige and proliferation of fiction (some 40,000 titles are said to have been published in this period), this age of the novel was in fact also an age of great poetry. Alfred Tennyson made his mark very early with *Poems, Chiefly Lyrical* (1830) and *Poems* (1832; dated 1833), publications that led some critics to hail him as the natural successor to Keats and Shelley. A decade later, in *Poems* (1842), Tennyson combined in two volumes the best of his early work with a second volume of more recent writing. The collection established him as the outstanding poet of the era.

Tennyson's *Poems*

In his early work Tennyson brought an exquisite lyric gift to late-Romantic subject matter. The result is a poetry that, for all its debt to Keats, anticipates the French Symbolists of the 1880s. The death of his friend and supporter Arthur Hallam in 1833, however, left him vulnerable to accusations from less sympathetic critics that this highly subjective verse was insufficiently engaged with the public issues of the day. The second volume of the *Poems* of 1842 contains two remarkable responses to this challenge. One is the dramatic monologue, a technique developed independently by both Tennyson and Browning in the 1830s and the greatest formal innovation in Victorian poetry. The other is the form that Tennyson called the English Idyl, in which he combined brilliant vignettes of contemporary landscape with relaxed debate.

In the major poems of his middle period Tennyson combined the larger scale required by his new ambitions with his original gift for the brief lyric by building long poems out of short ones. *In Memoriam* (1850) is an elegy for Arthur Hallam, formed by 133 individual lyrics. Eloquent, vivid, and ample, it is at the same time an acute pathological study of individual grief and the central Victorian statement of the problems posed by the decline of Christian faith. *Maud* (1855) assembles 27 lyric poems into a single dramatic monologue that disturbingly explores the psychology of violence.

Tennyson became poet laureate in 1850 and wrote some apt and memorable poems on patriotic themes. The chief work of his later period, however, was *Idylls of the King* (1859, revised 1885). An Arthurian epic, it offers a sombre vision of an idealistic community in decay. Some passages are brilliant, but even Tennyson's contemporaries found it on the whole oddly inhibited and lacking in intellectual substance.

G.K. Chesterton described Tennyson as "a suburban Virgil." The elegant Virgilian note was the last thing aimed at by his great contemporary Robert Browning. Browning's work was Germanic rather than Italianate, grotesque rather than idyllic, and colloquial rather than refined. The differences between Browning and Tennyson underline the creative diversity of the period.

Robert Browning and Elizabeth Barrett Browning. Deeply influenced by Shelley, Browning made two false starts. One was as a playwright, an ambition in which he persisted until 1846 and of which the one memorable product is *Pippa Passes* (1841). The other was as the late-Romantic poet of the confessional meditation *Pauline* (1833), the

closet drama *Paracelsus* (1835), and the difficult though innovatory narrative poem *Sordello* (1840).

Browning's *Dramatic Lyrics*

Browning found his individual and distinctively modern voice in 1842, with the volume *Dramatic Lyrics.* As the title suggests, it was a collection of dramatic monologues, among them "Porphyria's Lover," "Johannes Agricola in Meditation," and "My Last Duchess." The monologues make clear the radical originality of Browning's new manner: they involve the reader in sympathetic identification with the interior processes of criminal or unconventional minds, requiring active rather than merely passive engagement in the processes of moral judgment and self-discovery. More such monologues and some equally striking lyrics make up *Men and Women* (1855).

In 1846 Browning married Elizabeth Barrett. Though now remembered chiefly for her love poems *Sonnets from the Portuguese* (1850) and her experiment with the verse novel *Aurora Leigh* (1856; dated 1857), she was in her own lifetime far better known than her husband. Only with the publication of *Dramatis Personae* (1864) did Browning achieve the sort of fame that Tennyson had enjoyed for more than 20 years. The volume contains, in "Rabbi Ben Ezra," the most extreme statement of Browning's celebrated optimism. Hand in hand with this reassuring creed, however, go the skeptical intelligence and the sense of the grotesque displayed in such poems as "Caliban upon Setebos" and "Mr. Sludge, 'The Medium.' "

The Ring and the Book (1868–69) gives the dramatic monologue format unprecedented scope. Published in parts, like a Dickens novel, it tells a sordid murder story in a way that both explores moral issues and suggests the problematic nature of human knowledge. Browning's work after this date, though voluminous, is uneven.

Arnold and Clough. Matthew Arnold's first volume of verse, *The Strayed Reveller, and Other Poems* (1849), combined lyric grace with an acute sense of the dark philosophical landscape of the period. The title poem of his next collection, *Empedocles on Etna* (1852), is a sustained statement of the modern dilemma and a remarkable poetic embodiment of the process that Arnold called "the dialogue of the mind with itself." Arnold later suppressed this poem and attempted to write in a more impersonal manner. His greatest work ("Switzerland," "Dover Beach," "The Scholar-Gipsy") is, however, always elegiac in tone. In the 1860s he turned from verse to prose and became, with *Essays in Criticism* (1865), *Culture and Anarchy* (1869), and *Literature and Dogma* (1873), a lively and acute writer of literary, social, and religious criticism.

Arnold's friend Arthur Hugh Clough died young but managed, nonetheless, to produce three highly original poems. *The Bothie of Tober-na-Vuolich* (1848) is a narrative poem of modern life, written in hexameters. *Amours de Voyage* (1858) goes beyond this to the full-scale verse novel, using multiple internal narrators and vivid contemporary detail. *Dipsychus* (published posthumously in 1865 but not available in an unexpurgated version until 1951) is a remarkable closet drama that debates issues of belief and morality with a frankness, and a metrical liveliness, unequaled in Victorian verse.

EARLY VICTORIAN NONFICTIONAL PROSE

Carlyle's *French Revolution*

Carlyle may be said to have initiated Victorian literature with *Sartor Resartus.* He continued thereafter to have a powerful effect on its development. *The French Revolution* (1837), the book that made him famous, spoke very directly to this consciously postrevolutionary age. *On Heroes, Hero-Worship, and the Heroic in History* (1841) combined the Romantic idea of the genius with a further statement of the German transcendentalist philosophy, which Carlyle opposed to the influential doctrines of empiricism and utilitarianism. Carlyle's political writing, in *Chartism* (1839; dated 1840), *Past and Present* (1843), and the splenetic *Latter-day Pamphlets* (1850), inspired other writers to similar "prophetic" denunciations of laissez-faire economics and utilitarian ethics. The first importance of John Ruskin is as an art critic who, in *Modern Painters,* five volumes (1843–60), brought Romantic theory to the study of painting and forged an appropriate prose for its expression. But in *The Stones of Venice,* three volumes (1851–53), Ruskin took the political medievalism of Carlyle's *Past and Present* and gave it a poetic fullness and force. This imaginative engagement with social and economic problems continued into *Unto This Last* (1860), *The Crown of Wild Olive* (1866), and *Fors Clavigera* (1871–84). John Henry Newman was a poet, novelist, and theologian who wrote many of the tracts, published as *Tracts for the Times* (1833–41), that promoted the Oxford Movement in the Church of England. His subsequent religious development is memorably described in his *Apologia pro Vita Sua* (1864), one of the many great autobiographies of this introspective century.

LATE VICTORIAN LITERATURE

"The modern spirit," Matthew Arnold observed in 1865, "is now awake." In 1859 Charles Darwin had published *On the Origin of Species by Means of Natural Selection.* Historians, philosophers, and scientists were all beginning to apply the idea of evolution to new areas of study of the human experience. Traditional conceptions of man's nature and place in the world were, as a consequence, under threat. Walter Pater summed up the process, in 1866, by stating that "Modern thought is distinguished from ancient by its cultivation of the 'relative' spirit in place of the 'absolute.' "

The economic crisis of the 1840s was long past. But the fierce political debates that led first to the Second Reform Act of 1867 and then to the battles for the enfranchisement of women were accompanied by a deepening crisis of belief.

The novel. Late Victorian fiction may express doubts and uncertainties, but in aesthetic terms it displays a new sophistication and self-confidence. The American novelist Henry James wrote in 1884 that until recently the English novel had "had no air of having a theory, a conviction, a consciousness of itself behind it." Its acquisition of these things was due in no small part to Mary Ann Evans, better known as George Eliot. Initially a critic and translator, she was influenced, after the loss of her Christian faith, by the ideas of Ludwig Feuerbach and Auguste Comte. Her advanced intellectual interests combined with her sophisticated sense of the novel form to shape her remarkable fiction. Her early novels, *Adam Bede* (1859), *The Mill on the Floss* (1860), and *Silas Marner* (1861), are closely observed studies of English rural life that offer, at the same time, complex contemporary ideas and a subtle tracing of moral issues. Her masterpiece, *Middlemarch* (1871–72), is an unprecedentedly full study of the life of a provincial town, focused on the thwarted idealism of her two principal characters. George Eliot is a realist, but her realism involves a scientific analysis of the interior processes of social and personal existence.

George Eliot

Her fellow realist Anthony Trollope published his first novel in 1847 but only established his distinctive manner with *The Warden* (1855), the first of a series of six novels set in the fictional county of Barsetshire and completed in 1867. This sequence was followed by a further series, the six-volume Palliser group (1864–80), set in the world of British parliamentary politics. Trollope published an astonishing total of 47 novels, and his *Autobiography* (1883) is a uniquely candid account of the working life of a Victorian writer.

Trollope's series of novels

The third major novelist of the 1870s was George Meredith, who also worked as poet, journalist, and publisher's reader. His prose style is eccentric and his achievement uneven. His greatest work of fiction, *The Egoist* (1879), however, is an incisive comic novel that embodies the distinctive theory of the corrective and therapeutic powers of laughter expressed in his lecture "The Idea of Comedy" (1877).

This flowering of realist fiction was accompanied, perhaps inevitably, by a revival of its opposite, the romance. The 1860s produced a new subgenre, the sensation novel, seen at its best in the work of Wilkie Collins. Gothic novels and romances by Sheridan Le Fanu, Robert Louis Stevenson, William Morris, and Oscar Wilde; utopian fiction by Morris and Samuel Butler; and the early science fiction of H.G. Wells make it possible to speak of a full-scale romance revival.

Realism continued, however, to flourish, sometimes encouraged by the example of European Realist and Naturalist novelists. Both George Moore and George Gissing were influenced by Émile Zola, though both also reacted against him. The greatest novelist of this generation, however, was Thomas Hardy. His first published novel, *Desperate Remedies,* appeared in 1871 and was followed by 13 more before he abandoned prose to publish (in the 20th century) only poetry. His major fiction consists of the tragic novels of rural life, *The Mayor of Casterbridge* (1886), *Tess of the D'Urbervilles* (1891), and *Jude the Obscure* (1895). In these novels his brilliant evocation of the landscape and people of his fictional Wessex is combined with a sophisticated sense of "the ache of modernism."

Hardy's major fiction

Verse. The Pre-Raphaelite Brotherhood, formed in 1848 and unofficially reinforced a decade later, was founded as a group of painters but also functioned as a school of writers who linked the incipient Aestheticism of Keats and De Quincey to the Decadent movement of the fin de siècle. Dante Gabriel Rossetti collected his early writing in *Poems* (1870), a volume that led the critic Robert Buchanan to attack him as the leader of "The Fleshly School of Poetry." Rossetti combined some subtle treatments of contemporary life with a new kind of medievalism, seen also in *The Defence of Guenevere* (1858) by William Morris. The earnest political use of the Middle Ages found in Carlyle and Ruskin did not die out—Morris himself continued it and linked it, in the 1880s, with Marxism. But these writers also used medieval settings as a context that made possible an uninhibited treatment of sex and violence. The shocking subject matter and vivid imagery of Morris' first volume were further developed by Algernon Charles Swinburne, who, in *Atalanta in Calydon* (1865) and *Poems and Ballads* (1866), combined them with an intoxicating metrical power.

The carefully wrought religious poetry of Christina Rossetti is perhaps truer to the original, pious purposes of the Pre-Raphaelite Brotherhood. More interesting as a religious poet of this period, however, is Gerard Manley Hopkins, a Jesuit priest whose work was first collected as *Poems* in 1918, nearly 30 years after his death. Overpraised by modernist critics, who saw him as the sole great poet of the era, he was in fact an important minor talent and an ingenious technical innovator.

Hopkins' *Poems*

The 1890s witnessed a flowering of lyric verse, influenced intellectually by the critic and novelist Walter Pater and formally by contemporary French practice. Such writing was widely attacked as "decadent" for its improper subject matter and its consciously amoral doctrine of "art for art's sake." This stress upon artifice and the freedom of art from conventional moral constraints went hand in hand, however, with an exquisite craftsmanship and a devotion to intense emotional and sensory effects. Outstanding among the numerous poets publishing in the final decade of the century were John Davidson, Arthur Symons, Francis Thompson, Ernest Dowson, Lionel Johnson, and A.E. Housman. In *The Symbolist Movement in Literature* (1899) Symons suggested the links between this writing and European Symbolism and Impressionism. Thompson provides a vivid example of the way in which a decadent manner could, paradoxically, be combined with fierce religious enthusiasm. A rather different note was struck by Rudyard Kipling, who combined polemical force and sharp observation (particularly of colonial experience) with a remarkable metrical vigour.

THE VICTORIAN THEATRE

Early Victorian drama was a popular art form, appealing to an uneducated audience that demanded emotional excitement rather than intellectual subtlety. Vivacious melodramas did not, however, hold exclusive possession of the stage. The mid-century saw lively comedies by Dion Boucicault and Tom Taylor. In the 1860s T.W. Robertson pioneered a new realist drama, an achievement later celebrated by Arthur Wing Pinero in his charming sentimental comedy *Trelawny of the "Wells"* (1898). The 1890s were, however, the outstanding decade of dramatic innovation. Oscar Wilde crowned his brief career as a playwright with one of the few great high comedies in English, *The Importance of Being Earnest* (1895). At the same time the influence of Henrik Ibsen was helping to produce a new genre of serious "problem plays," such as Pinero's *Second Mrs. Tanqueray* (1893). J.T. Grein founded the Independent Theatre in 1891 to foster such work and staged there the first plays of George Bernard Shaw and translations of Ibsen.

VICTORIAN LITERARY COMEDY

Victorian literature began with such humorous books as *Sartor Resartus* and *The Pickwick Papers.* Despite the crisis of faith, the "Condition of England" question, and the ache of modernism, this note was sustained throughout the century. The comic novels of Dickens and Thackeray; the squibs, sketches, and light verse of Thomas Hood and Douglas Jerrold; the nonsense of Edward Lear and Lewis Carroll; and the humorous light fiction of Jerome K. Jerome and George Grossmith and his brother Weedon Grossmith are proof that this age, so often remembered for its gloomy rectitude, may in fact have been the greatest era of comic writing in English literature. (N.Sh.)

"Modern" English literature: the 20th century

FROM 1900 TO 1945

The Edwardians. The 20th century opened with great hope but also with some apprehension, for the new century marked the onset of a new millennium. For many, mankind was entering upon an unprecedented era. H.G. Wells's utopian studies, the aptly titled *Anticipations of the Reaction of Mechanical and Scientific Progress upon Human Life and Thought* (1901) and *A Modern Utopia* (1905), both captured and qualified this optimistic mood and gave expression to a common conviction that science and technology would transform the world in the century ahead. To achieve such transformation, outmoded institutions and ideals had to be replaced by ones more suited to the growth and liberation of the human spirit. The death of Queen Victoria in 1901 and the accession of Edward VII seemed to confirm that a franker, less inhibited era had begun.

Many writers of the Edwardian period, drawing widely upon the realistic and naturalistic conventions of the 19th century (upon Ibsen in drama and Balzac, Turgenev, Flaubert, Zola, Eliot, and Dickens in fiction) and in tune with the anti-Aestheticism unleashed by the trial of the archetypal Aesthete, Oscar Wilde, saw their task in the new century to be an unashamedly didactic one. In a series of wittily iconoclastic plays, of which *Man and Superman* (performed 1905, published 1903) and *Major Barbara* (performed 1905, published 1907) are the most substantial, George Bernard Shaw turned the Edwardian theatre into an arena for debate upon the principal concerns of the day: the question of political organization, the morality of armaments and war, the function of class and of the professions, the validity of the family and of marriage, and the issue of female emancipation. Nor was he alone in this, even if he was alone in the brilliance of his comedy. John Galsworthy made use of the theatre in *Strife* (1909) to explore the conflict between capital and labour, and in *Justice* (1910) he lent his support to reform of the penal system, while Harley Granville-Barker, whose revolutionary approach to stage direction did much to change theatrical production in the period, dissected in *The Voysey Inheritance* (performed 1905, published 1909) and *Waste* (performed 1907, published 1909) the hypocrisies and deceit of upper-class and professional life.

Didactic literature

Many Edwardian novelists were similarly eager to explore the shortcomings of English social life. Wells—in *Love and Mr. Lewisham* (1900); *Kipps* (1905); *Ann Veronica* (1909), his pro-suffragette novel; and *The History of Mr. Polly* (1910)—captured the frustrations of lower- and middle-class existence, even though he relieved his accounts with many comic touches. In *Anna of the Five Towns* (1902) Arnold Bennett detailed the constrictions of provincial life among the self-made business classes in the area of England known as the Potteries; in *The Man of Property* (1906), the first volume of *The Forsyte Saga,* Galsworthy described the destructive possessiveness of the professional

bourgeoisie; and in *Where Angels Fear to Tread* (1905) and *The Longest Journey* (1907) E.M. Forster portrayed with irony the insensitivity, self-repression, and philistinism of the English middle classes.

These novelists, however, wrote more memorably when they allowed themselves a larger perspective. In *The Old Wives' Tale* (1908) Bennett showed the destructive effects of time on the lives of individuals and communities and evoked a quality of pathos that he never matched in his other fiction; in *Tono-Bungay* (1909) Wells showed the ominous consequences of the uncontrolled developments taking place within a British society still dependent upon the institutions of a long-defunct landed aristocracy; and in *Howards End* (1910) Forster showed how little the rootless and self-important world of contemporary commerce cared for the more rooted world of culture, although he acknowledged that commerce was a necessary evil. Nevertheless, even as they perceived the difficulties of the present, most Edwardian novelists, like their counterparts in the theatre, held firmly to the belief not only that constructive change was possible but also that this change could in some measure be advanced by their writing.

Other writers, including Thomas Hardy and Rudyard Kipling, who had established their reputations during the previous century, and Hilaire Belloc, G.K. Chesterton, and Edward Thomas, who established their reputations in the first decade of the new century, were less confident about the future and sought to revive the traditional forms—the ballad, the narrative poem, the satire, the fantasy, the topographical poem, and the essay—that in their view preserved traditional sentiments and perceptions. The revival of traditional forms in the late 19th and early 20th century was not a unique event. There have been many such revivals during the 20th century, and the traditional poetry of A.E. Housman (whose book *A Shropshire Lad*, originally published in 1896, enjoyed huge popular success during World War I), Walter de la Mare, John Masefield, Robert Graves, and Edmund Blunden represents an important and often neglected strand of English literature in the first half of the century.

Revival of traditional forms

The most significant writing of the period, traditionalist or modern, was inspired by neither hope nor apprehension but by bleaker feelings that the new century would witness the collapse of a whole civilization. The new century had begun with Great Britain involved in the South African War (the Boer War; 1899–1902), and it seemed to some that the British Empire was as doomed to destruction, both from within and from without, as had been the Roman Empire. In his poems on the South African War, Hardy (whose achievement as a poet in the 20th century rivaled his achievement as a novelist in the 19th) questioned simply and sardonically the human cost of empire building and established a tone and style that many British poets were to use in the course of the century, while Kipling, who had done much to engender pride in empire, began to speak in his verse and short stories of the burden of empire and the tribulations it would bring.

James's sense of an imperial civilization in decline

No one captured the sense of an imperial civilization in decline more fully or subtly than the expatriate American novelist Henry James. In *The Portrait of a Lady* (1881) he had briefly anatomized the fatal loss of energy of the English ruling class and in *The Princess Casamassima* (1886) had described more directly the various instabilities that threatened its paternalistic rule. He did so with regret: the patrician American admired in the English upper class its sense of moral obligation to the community. By the turn of the century, however, he had noted a disturbing change. In *The Spoils of Poynton* (1897) and *What Maisie Knew* (1897) members of the upper class no longer seem troubled by the means adopted to achieve their morally dubious ends. Great Britain had become indistinguishable from the other nations of the Old World, in which an ugly rapacity had never been far from the surface. James's dismay at this condition gave to his subtle and compressed late fiction, *The Wings of the Dove* (1902), *The Ambassadors* (1903), and *The Golden Bowl* (1904), much of its gravity and air of disenchantment.

James's awareness of crisis affected the very form and style of his writing, for he was no longer assured that the world about which he wrote was either coherent in itself or unambiguously intelligible to its inhabitants. His fiction still presented characters within an identifiable social world, but he found his characters and their world increasingly elusive and enigmatic and his own grasp upon them, as he made clear in *The Sacred Fount* (1901), the questionable consequence of artistic will.

Another expatriate novelist, Joseph Conrad (pseudonym of Józef Teodor Konrad Korzeniowski, born in the Ukraine of Polish parents), shared James's sense of crisis but attributed it less to the decline of a specific civilization than to the failings of mankind itself. Man was a solitary, romantic creature of will who at any cost imposed his meaning upon the world because he could not endure a world that did not reflect his central place within it. In *Almayer's Folly* (1895) and *Lord Jim* (1900) he had seemed to sympathize with this predicament; but in "Heart of Darkness" (1902), *Nostromo* (1904), *The Secret Agent* (1907), and *Under Western Eyes* (1911) he detailed such imposition, and the psychological pathologies he increasingly associated with it, without sympathy. He did so as a philosophical novelist whose concern with the mocking limits of human knowledge affected not only the content of his fiction but also its very structure. His writing itself is marked by gaps in the narrative, by narrators who do not fully grasp the significance of the events they are retelling, and by characters who are unable to make themselves understood. James and Conrad used many of the conventions of 19th-century realism but transformed them to express what are considered to be peculiarly 20th-century preoccupations and anxieties.

The modernist revolution. *Anglo-American modernism: Pound, Lewis, Lawrence, and Eliot.* From 1908 to 1914 there was a remarkably productive period of innovation and experiment as novelists and poets undertook, in anthologies and magazines, to challenge the literary conventions not just of the recent past but of the entire Post-Romantic era. For a brief moment, London, which up to that point had been culturally one of the dullest of the European capitals, boasted an avant-garde to rival those of Paris, Vienna, and Berlin, even if its leading personality, Ezra Pound, and many of its most notable figures were American.

London's avant-garde

The spirit of modernism—a radical and utopian spirit stimulated by new ideas in anthropology, psychology, philosophy, political theory, and psychoanalysis—was in the air, expressed rather mutedly by the pastoral and often anti-modern poets of the Georgian movement (1912–22) and more authentically by the English and American poets of the Imagist movement, to which Pound first drew attention in *Ripostes* (1912), a volume of his own poetry, and in *Des Imagistes* (1914), an anthology. Prominent among the Imagists were the English poets T.E. Hulme, F.S. Flint, and Richard Aldington and the Americans Hilda Doolittle (H.D.) and Amy Lowell.

Reacting against what they considered to be an exhausted poetic tradition, the Imagists wanted to refine the language of poetry in order to make it a vehicle not for pastoral sentiment or imperialistic rhetoric but for the exact description and evocation of mood. To this end they experimented with free or irregular verse and made the image their principal instrument. In contrast to the leisurely Georgians, they worked with brief and economical forms.

Meanwhile, painters and sculptors, grouped together by the painter and writer Wyndham Lewis under the banner of vorticism, combined the abstract art of the Cubists with the example of the Italian Futurists who conveyed in their painting, sculpture, and literature the new sensations of movement and scale associated with such modern developments as automobiles and airplanes. With the typographically arresting *Blast: Review of the Great English Vortex* (two editions, 1914 and 1915) vorticism found its polemical mouthpiece and in its editor, Wyndham Lewis, its most active propagandist and accomplished literary exponent. His experimental play *Enemy of the Stars*, published in *Blast* in 1914, and his experimental novel *Tarr* (1918) can still surprise with their violent exuberance.

World War I brought this first period of the modernist revolution to an end and, while not destroying its radical

and utopian impulse, made the Anglo-American modernists all too aware of the gulf between their ideals and the chaos of the present. Novelists and poets parodied received forms and styles, in their view made redundant by the immensity and horror of the war, but, as can be seen most clearly in Pound's angry and satirical *Hugh Selwyn Mauberley* (1920), with a note of anguish and with the wish that writers might again make form and style the bearers of authentic meanings.

Lawrence and Eliot on the sickness of modern civilization

In his two most innovative novels, *The Rainbow* (1915) and *Women in Love* (1920), D.H. Lawrence traced the sickness of modern civilization—a civilization in his view only too eager to participate in the mass slaughter of the war—to the effects of industrialization upon the human psyche. Yet as he rejected the conventions of the fictional tradition, which he had used to brilliant effect in his deeply-felt autobiographical novel of working-class family life, *Sons and Lovers* (1913), he drew upon myth and symbol to hold out the hope that individual and collective rebirth could come through human intensity and passion.

On the other hand, the poet and playwright T.S. Eliot, another American resident in London, in his most innovative poetry, *Prufrock and Other Observations* (1917) and *The Waste Land* (1922), traced the sickness of modern civilization—a civilization that, on the evidence of the war, preferred death or death-in-life to life—to the spiritual emptiness and rootlessness of modern existence. As he rejected the conventions of the poetic tradition, Eliot, like Lawrence, drew upon myth and symbol to hold out the hope of individual and collective rebirth, but he differed sharply from Lawrence by supposing that rebirth could come through self-denial and self-abnegation. Even so, their satirical intensity, no less than the seriousness and scope of their analyses of the failings of a civilization that had voluntarily entered upon the first World War, ensured that Lawrence and Eliot became the leading and most authoritative figures of Anglo-American modernism in England in the whole of the postwar period.

During the 1920s, Lawrence (who left England in 1919) and Eliot began to develop viewpoints at odds with the reputations they had established through their early work. In *Kangaroo* (1923) and *The Plumed Serpent* (1926) Lawrence revealed the attraction to him of charismatic, masculine leadership, while in *For Lancelot Andrewes: Essays on Style and Order* (1928) Eliot (whose influence as a literary critic now rivaled his influence as a poet) announced that he was a "classicist in literature, royalist in politics and anglo-catholic in religion" and committed himself to hierarchy and order. Elitist and paternalistic, they did not, however, adopt the extreme positions of Pound (who left England in 1920 and settled permanently in Italy in 1925) or Lewis. Drawing upon the ideas of the left and of the right, Pound and Lewis dismissed democracy as a sham and argued that economic and ideological manipulation was the dominant factor. For some the antidemocratic views of the Anglo-American modernists simply made explicit the reactionary tendencies inherent in the movement from its beginning; for others they came from a tragic loss of balance occasioned by World War I. This issue is a complex one, and judgments upon the literary merit and political status of Pound's ambitious but immensely difficult imagist epic *The Cantos* (1917–70) and Lewis' powerful sequence of politico-theological novels *The Human Age* (*The Childermass,* 1928; *Monstre Gai* and *Malign Fiesta,* both 1955) are sharply divided.

Celtic modernism: Yeats, Joyce, Jones, and MacDiarmid. Pound, Lewis, Lawrence, and Eliot were the principal figures of Anglo-American modernism, but important contributions also were made by the Irish poet and playwright William Butler Yeats and the Irish novelist James Joyce. By virtue of nationality, residence, and, in Yeats's case, an unjust reputation as a poet still steeped in Celtic mythology, they had less immediate impact upon the British literary intelligentsia in the late 1910s and early 1920s than Pound, Lewis, Lawrence, and Eliot, although by the mid-1920s their influence had become direct and substantial. Many contemporary critics argue that Yeats's work as a poet and Joyce's work as a novelist are the most important modernist achievements of the period.

In his early verse and drama Yeats, who had been influenced as a young man by the Romantic and Pre-Raphaelite movements, evoked a legendary and supernatural Ireland in language that was often vague and grandiloquent. As an adherent of the cause of Irish nationalism he had hoped to instill pride in the Irish past. The poetry of *The Green Helmet* (1910) and *Responsibilities* (1914), however, was marked not only by a more concrete and colloquial style but also by a growing isolation from the nationalist movement, for Yeats celebrated an aristocratic Ireland epitomized for him by the family and country house of his friend and patron, Lady Gregory.

The grandeur of his mature reflective poetry in *The Wild Swans at Coole* (1917), *Michael Robartes and the Dancer* (1921), *The Tower* (1928), and *The Winding Stair* (1929) derived in large measure from the way in which (caught up by the violent discords of contemporary Irish history) he accepted the fact that his idealized Ireland was illusory. At its best his mature style combined passion and precision with powerful symbol, strong rhythm, and lucid diction; and even though his poetry often touched upon public themes, he never ceased to reflect upon the Romantic themes of creativity, selfhood, and the individual's relationship to nature, time, and history.

Joyce, who spent his adult life on the continent of Europe, expressed in his fiction his sense of the limits and possibilities of the Ireland he had left behind. In his collection of short stories, *Dubliners* (1914), and his largely autobiographical novel, *A Portrait of the Artist as a Young Man* (1916), he described in fiction at once realist and symbolist the individual cost of the sexual and imaginative oppressiveness of life in Ireland. As if by provocative contrast, his panoramic novel of urban life, *Ulysses* (1922), was sexually frank and imaginatively profuse. (Copies of the first edition were burned by the New York postal authorities, and British customs officials seized the second edition in 1923.) Employing extraordinary formal and linguistic inventiveness, including the so-called stream-of-consciousness method, Joyce depicted the experiences and the fantasies of various men and women in Dublin on a summer's day in June 1904. Yet his purpose was not simply documentary, for he drew upon an encyclopaedic range of European literature to stress the rich universality of life buried beneath the provincialism of pre-independence Dublin, still in 1904 a city within the British Empire. In his even more experimental *Finnegans Wake* (1939), extracts of which had already appeared as *Work in Progress* from 1928 to 1937, Joyce's commitment to cultural universality became absolute. By means of a strange, polyglot idiom of puns and portmanteau words he not only explored the relationship between the conscious and the unconscious but also suggested that the languages and myths of Ireland were interwoven with the languages and myths of many other cultures.

Joyce's inventive fiction

The example of Joyce's experimentalism was followed by the Anglo-Welsh poet David Jones and by the Scottish poet Hugh MacDiarmid (pseudonym of Christopher Murray Grieve). Whereas Jones concerned himself, in his complex and allusive poetry and prose, with the Celtic, Saxon, Roman, and Christian roots of Great Britain, MacDiarmid sought not only to recover what he considered to be an authentically Scottish culture but also to establish, as in his *In Memoriam James Joyce* (1955), the truly cosmopolitan nature of Celtic consciousness and achievement. MacDiarmid's masterpiece in the vernacular, *A Drunk Man Looks at the Thistle* (1926), helped to inspire the Scottish renaissance of the 1920s and '30s.

The literature of World War I and the interwar period. The impact of World War I upon the Anglo-American modernists has been noted. In addition the war brought a variety of responses from the more traditionalist writers, predominantly poets, who saw action. Rupert Brooke caught the idealism of the opening months of the war (and died in service); Siegfried Sassoon and Ivor Gurney caught the mounting anger and sense of waste as the war continued; and Isaac Rosenberg (perhaps the most original of the war poets), Wilfrid Owen, and Edmund Blunden not only caught the comradely compassion of the trenches but also addressed themselves to the larger moral

perplexities raised by the war (Rosenberg and Owen were killed in action).

Cynicism in the postwar years

It was not until the 1930s, however, that much of this poetry became widely known. In the wake of the war the dominant tone, at once cynical and bewildered, was set by Aldous Huxley's satirical novel *Crome Yellow* (1921). Drawing upon Lawrence and Eliot, he concerned himself in his novels of ideas—*Antic Hay* (1923), *Those Barren Leaves* (1925), and *Point Counter Point* (1928)—with the fate of the individual in rootless modernity. His pessimistic vision found its most complete expression in the 1930s, however, in his most famous and inventive novel, the anti-utopian fantasy *Brave New World* (1932), and his account of the anxieties of middle-class intellectuals of the period, *Eyeless in Gaza* (1936).

Huxley's frank and disillusioned manner was echoed by the poet Robert Graves in his autobiography, *Good-bye to All That* (1929), and by the poet Richard Aldington in his *Death of a Hero* (1929), a semiautobiographical novel of prewar bohemian London and the trenches. Exceptions to this dominant mood were found among writers too old to consider themselves, as did Graves and Aldington, members of a betrayed generation. In *A Passage to India* (1924) E.M. Forster examined the quest for and failure of human understanding among various ethnic and social groups in India under British rule. In *Parade's End* (1950; comprising *Some Do Not,* 1924; *No More Parades,* 1925; *A Man Could Stand Up,* 1926; and *Last Post,* 1928) Ford Madox Ford, with an obvious debt to James and Conrad, examined the demise of aristocratic England in the course of the war, exploring on a larger scale the themes he had treated with brilliant economy in his short novel *The Good Soldier* (1915). And in *Wolf Solent* (1929) and *A Glastonbury Romance* (1932), John Cowper Powys developed an eccentric and highly erotic mysticism.

These were, however, writers of an earlier, more confident era. A younger and more contemporary voice belonged to members of the Bloomsbury group. Setting themselves against the humbug and hypocrisy that, they believed, had marked their parents' generation in upper-class England, they aimed to be uncompromisingly honest in personal and artistic life. In Lytton Strachey's iconoclastic biographical study *Eminent Victorians* (1918) this amounted to little more than amusing irreverence, even though Strachey had a profound effect upon the writing of biography; but in the fiction of Virginia Woolf the rewards of this outlook were both profound and moving.

Virginia Woolf's fiction and essays

In short stories and novels of great delicacy and lyrical power she set out to portray the limitations of the self, caught as it is in time, and suggested that these could be transcended, if only momentarily, by engagement with another self, a place, or a work of art. This preoccupation not only charged the act of reading and writing with unusual significance but also produced, in *To the Lighthouse* (1927), *The Waves* (1931)—perhaps her most inventive and complex novel—and *Between the Acts* (1941), her most sombre and moving work, some of the most daring fiction produced in the 20th century.

Woolf believed that her viewpoint offered an alternative to the destructive egotism of the masculine mind, an egotism that had found its outlet in World War I, but she did not consider this viewpoint, as she made clear in her essay *A Room of One's Own* (1929), to be the unique possession of women. In her fiction she presented men who possessed what she held to be feminine characteristics, a regard for others and an awareness of the multiplicity of experience; but she remained pessimistic about women gaining positions of influence, even though she set out the desirability of this in her feminist study *Three Guineas* (1938). Together with Joyce, who greatly influenced her *Mrs. Dalloway* (1925), Woolf transformed the treatment of subjectivity, time, and history in fiction and helped create a feeling among her contemporaries that traditional forms of fiction—with their frequent indifference to the mysterious and inchoate inner life of characters—were no longer adequate. Her eminence as a literary critic and essayist did much to foster an interest in the writing of other significant women novelists, such as Katherine Mansfield and Dorothy Richardson.

The 1930s. World War I created a profound sense of crisis in English culture, and this became even more intense with the worldwide economic collapse of the late 1920s and early '30s, the rise of Fascism, the Spanish Civil War (1936–39), and the approach of another full-scale conflict in Europe. It is not surprising, therefore, that much of the writing of the 1930s was bleak and pessimistic: even Evelyn Waugh's sharp and amusing satire on contemporary England, *Vile Bodies* (1930), ended with another, more disastrous war.

Divisions of class and the burden of sexual repression became common and interrelated themes in the fiction of the 1930s, a fiction that largely neglected the modernist revolution in technique of the 1920s and returned to the realist modes of the first decade of the century. In *A Scots Quair* (*Sunset Song,* 1932; *Cloud Howe,* 1933; and *Grey Granite,* 1934) the novelist Lewis Grassic Gibbon (pseudonym of James Leslie Mitchell) gives a panoramic account of Scottish rural and working-class life. The work resembles Lawrence's novel *The Rainbow* in its historical sweep and intensity of vision. Walter Greenwood's *Love on the Dole* (1933) is a bleak record, in the manner of Bennett, of the economic depression in a northern working-class community; and Graham Greene's *It's a Battlefield* (1934) and *Brighton Rock* (1938) are desolate studies, in the manner of Conrad, of the loneliness and guilt of men and women trapped in a contemporary England of conflict and decay. *A Clergyman's Daughter* (1935) and *Keep the Aspidistra Flying* (1936), by George Orwell, are evocations, in the manner of Wells and, in the latter case unsuccessfully, of Joyce, of contemporary lower middle-class existence, and *The Road to Wigan Pier* (1937) is a report of northern working-class mores. Elizabeth Bowen's *Death of the Heart* (1938) is a sardonic analysis, in the manner of James, of contemporary upper-class values.

Realist modes of the 1930s

Yet the most interesting writing of the decade grew out of the determination to supplement the diagnosis of class division and sexual repression with their cure. It was no accident that the poetry of W.H. Auden and his Oxford contemporaries, C. Day-Lewis, Louis MacNeice, and Stephen (later Sir Stephen) Spender, became quickly identified as the authentic voice of the new generation, for it matched despair with defiance. These self-styled prophets of a new world envisaged freedom from the bourgeois order being achieved in various ways. For Day-Lewis and Spender technology held out particular promise. This, allied to Marxist precepts, would in their view bring an end to poverty and the suffering it caused. For Auden especially, sexual repression was the enemy, and here the writings of Sigmund Freud and D.H. Lawrence were valuable. Whatever their individual preoccupations, these poets produced in the very play of their poetry, with its mastery of different genres, its rapid shifts of tone and mood, and its strange juxtapositions of the colloquial and esoteric, a blend of seriousness and high spirits irresistible to their peers.

The adventurousness of the new generation was shown, in part, by its love of travel (as in Christopher Isherwood's novels *Mr. Norris Changes Trains* [1935] and *Goodbye to Berlin* [1939], which reflect his experiences of postwar Germany); in part by its readiness for political involvement; and in part by its openness to the writing of the avant-garde of the Continent. The verse dramas coauthored by Auden and Isherwood, of which *The Ascent of F6* (1936) is the most notable, owed much to Bertolt Brecht; the political parables of Rex Warner, of which *The Aerodrome* (1941) is the most accomplished, owed much to Franz Kafka; and the complex and often obscure poetry of David Gascoyne and Dylan Thomas owed much to the Surrealists. Even so, Yeats's mature poetry and Eliot's *Waste Land,* with its parodies, its satirical edge, its multiplicity of styles, and its quest for spiritual renewal, provided the most significant models and inspiration for the young writers of the period. On the whole, however, despite the breadth, diversity, and liveliness of the writing of the 1930s, the decade was not one of great originality or innovation but rather one of imitation and emulation.

The literature of World War II (1939–45). The outbreak of war in 1939, as in 1914, brought to an end an era

of great intellectual and creative exuberance. Individuals were dispersed; the rationing of paper affected the production of magazines and books; the poem and the short story, convenient forms for men under arms, became the favoured means of literary expression. It was hardly a time for new beginnings, although the poets of the New Apocalypse movement produced three anthologies (1940–45) inspired by neo-Romantic anarchism. No important new novelists or playwrights appeared, and only three new poets (all of whom died on active service) showed promise: Alun Lewis, Sidney Keyes, and Keith Douglas, the most gifted and distinctive, whose eerily detached accounts of the battlefield revealed a poet of potential greatness.

Eliot's *Four Quartets*

It was a poet of an earlier generation, T.S. Eliot, who produced in his *Four Quartets* (1935–42; published as a whole, 1943) the masterpiece of the war. Reflecting upon language, time, and history, he searched, in the three quartets written during the war, for moral and religious significance in the midst of destruction and strove to counter the spirit of nationalism inevitably present in a nation at war. The creativity that had seemed to end with the tortured religious poetry and verse drama of the 1920s and '30s had a rich and extraordinary late flowering as Eliot concerned himself, on the scale of *The Waste Land* but in a very different manner and mood, with the well-being of the society in which he lived. (H.A.Da.)

LITERATURE AFTER 1945

Postwar poetry and prose. When World War II ended the prevailing mood in literature was one of disillusionment, exhaustion, and sterility. Joyce, Yeats, and Virginia Woolf were all dead, and to George Orwell, writing in 1945, the younger English writers seemed like "fleas hopping among the ruins of a civilisation." Cyril Connolly, whose magazine *Horizon* had done so much to keep literary culture alive during the war, later remarked that, far from producing rapid change, the "effect of wars and catastrophes is to slow up the movement (so much more profound) of the human spirit." The novelist William Golding felt that the horrors of the war had beggared description: "We have discovered a limit to literature."

But if this sense of gloom and enervation was common to a wide range of writers at the end of the war, the responses it evoked were, as one might expect, various. For some writers the social and political unease of the 1930s, culminating in the experience of war, put an end to the further possibilities of modernism and created a hunger for more traditional values: T.S. Eliot, Dame Edith Sitwell, Evelyn Waugh, W.H. Auden, and Graham Greene turned increasingly to Christianity, while Aldous Huxley and Christopher Isherwood dabbled in mysticism and Eastern philosophy to the detriment of their later fiction. Alongside these spiritual hungers can be detected a certain restlessness typified by the later novels of Greene and by the fact that so many important writers—Huxley, Robert Graves, Auden, Isherwood—chose to live abroad.

Orwell. By contrast George Orwell stayed close to home, deriving his inspiration directly from the mood of Britain in the '40s. *Animal Farm* (1945) confronted the unpalatable truth that the victory over Fascism would in some respects unwittingly aid the advance of totalitarianism, while *Nineteen Eighty-four* (1949) warns of the dangers to the individual of encroaching collectivism. In these last, bleak fables Orwell attempted to make an art of political writing in the tradition of Swift and Defoe.

Eliot's verse dramas

Eliot. At the opposite end of the political spectrum was Eliot. After the *Four Quartets* he produced no more poetry, but he remained influential as a critic and publisher. His main creative effort now went into the theatre. Unlike his earlier poetic tragedies, his three postwar comedies (*The Cocktail Party,* 1950; *The Confidential Clerk,* 1954; and *The Elder Statesman,* 1959) constitute an attempt to write verse plays within the somewhat tired conventions of the commercial theatre. In consequence they now seem dated, although their intelligence and linguistic verve set them apart from the humdrum, if solidly crafted, drama of the period, typified at its best by plays such as J.B. Priestley's *An Inspector Calls* (1945).

Auden and Graves. With Eliot silent, the best poetry in the decade after the war was written by Auden and Graves, both still at the height of their powers. The brilliance and sharpness of Auden's earlier manner never deserted him, but he now began to develop the complex yet easily vernacular mixture of Horatian and Christian moralizing that formed the basis of his later style. In spite of his physical absence, the civilization and restrained charm of Auden's later voice, with its insistence on the private virtues, was to be a major and enduring influence on postwar English poetry. The influence of Graves was less direct but almost as important. Graves lived on Majorca like Prospero on his island, cultivating his magical powers as a poet unaffected by any literary fashion or movement—meanwhile supporting himself with a stream of highly competent popular novels and scholarly writings. A minor poet able to stand comparison with Auden and Graves is William Empson, whose *Collected Poems* (1949) marked the end of his slender poetic output. The power of his poetry is to convey moral toughness and moral nicety with an obscure passion that resists paraphrase or definition. The verse of John Betjeman is highly accessible, and he was undoubtedly the most popular poet laureate (1972–84) since Tennyson. His adherence to and mastery of traditional forms, his apparent simplicity, and his lightness of touch meant that his reputation was always uncertain; indeed, no poet of his quality has been more condescended to by the critics. But Betjeman was an innovator, not a sentimentalist; his sensibility was a modern one, keenly alive to the absurd. His nostalgia was built upon an exact sense of time and place and an instinct for the genuine.

The virtues of all these poets were more generally perceived in the 1950s, however, than in the period immediately after the war, when the dominant strain in English poetry ran counter to the cool irony and social observation of the '30s. There developed instead, for reasons partly connected with the war, an apocalyptic, neo-Romantic, and rhetorical tone in poetry with much emphasis placed on surrealism and the status of myth. The unruly deity of this cult was Dylan Thomas, its high priestess Dame Edith Sitwell, and its acolytes David Gascoyne and George Barker. Thomas was a wayward genius, a phrasemaker of power and originality who could seldom concentrate his thoughts through a whole poem, and whose effects, in consequence, are wearyingly random.

Waugh and his generation. A very different reaction to the war was perhaps the most lasting and also served as a necessary preface to many of the finest achievements in postwar literature. Very early in the war Evelyn Waugh predicted that "its chief use would be to cure artists of the illusion that they were men of action." Waugh's own honourable, if frustrating, wartime career was to confirm this view: "I don't want to influence opinions or events, or expose humbug or anything of that kind. I don't want to be of service to anyone or anything. I simply want to do my work as an artist." *Brideshead Revisited* (1945), Waugh's first attempt at the "magnum opus" that would widen the perspective of his prewar novels and tackle the operation of grace in human affairs, was, as he later acknowledged, deeply flawed by nostalgia and sentimentality. He was able to rework the theme completely in his late masterpiece *Sword of Honour* (1965; published separately as *Men at Arms,* 1952; *Officers and Gentlemen,* 1955; and *Unconditional Surrender,* 1961), bringing to the task new resources of humility and creative cunning. Passionate dedication to his art also enriches *The Ordeal of Gilbert Pinfold* (1957), a genuinely comic novel shot through with the sombre poignancy of Waugh's mature self-appraisal.

The other novelist of Waugh's generation to achieve a comparable stature in the postwar period was Anthony Powell. Powell's five prewar novels, never as popular as Waugh's, nonetheless show the emergence of a distinctive talent that was elaborated in his 12-novel sequence "A Dance to the Music of Time" (1951–75). Powell and Waugh are the master stylists of a generation that succeeded, in Powell's words, in "throwing overboard a good deal of Edwardian débris." They took hints about dialogue and presentation from the novels of Ronald Firbank and Dame Ivy Compton-Burnett, as well as the American writers Ernest Hemingway and e.e. cummings, and they both

were preoccupied with the "naturalist" problem of finding literary means to convey the essence of what people sound like, instead of merely reproducing what they say. But Powell is a very different novelist from Waugh. Much of Waugh's imaginative energy derives from a certain blinkered quality, an instinct for what can safely be left out, so that his themes and characters are sculpted with a hard and brilliant edge. Powell is a much more open and inclusive writer. His involvement with his characters is both more detached and more curious, he is interested in subtler distinctions and capable of a more genuine tenderness and lyricism, he is laconic but not cynical, and his wit, though quieter, is more deeply suffused through his work.

Two other novelists of this period were Henry Green and Joyce Cary. Green's novels are intriguingly empty at the centre; he has a viewpoint but no point of view. He is supremely a novelist of linguistic atmosphere. Each of his books takes a small subject (no two are similar) and transforms it into a self-sufficient world whose consistency is entirely made up of Green's delicate, sometimes whimsical, patterning of language and symbol. The technique is seen at its best in *Loving* (1945), a tenderly menacing study of servants and their masters in an Irish country house during the war. By contrast Cary, who began writing only after a career spent in colonial administration, is a gigantic and protean novelist of powerfully Protestant conviction expressed in his two trilogies (*Herself Surprised,* 1941; *To Be a Pilgrim,* 1942; and *The Horse's Mouth,* 1944; and *Prisoner of Grace,* 1952; *Except the Lord,* 1953; and *Not Honour More,* 1955).

Most of the writers so far considered were born well before World War I, but it would be misleading to suggest that no younger talents had yet emerged. Several good poets in their early 30s stood aside from the fashionable apocalyptic vein. R.S. Thomas, an admirer of the Metaphysical poet George Herbert, wrote spare, often bitter, lyrics arising from his experience as an Anglican priest in Wales. Roy Fuller was a poet of wry and modest exactitudes. The skill of Lawrence Durrell's mandarin and soberly reflective poetry has been regrettably eclipsed by the later success of his baroque and overwritten fiction. Two novelists of a distinctly unusual flavour were Malcolm Lowry (*Under the Volcano,* 1947) and Jocelyn Brooke (*The Military Orchid,* 1948; *A Mine of Serpents,* 1949; *The Goose Cathedral,* 1950; *The Dog at Clambercrown,* 1955). The most important new voice to be heard in the decade was Philip Larkin. Though *The North Ship* (1945) only hints at his development as a poet, it and his two novels (*Jill,* 1946; *A Girl in Winter,* 1947) were hopeful auguries for the new writing of the 1950s.

The 1950s and after. The 1950s are often conveniently categorized as the period in which the welfare state wrought great changes in British society, the time of the Angry Young Men, when in literature, as in other areas, class barriers were broken down. Such accounts have only a very limited use, for the period was one of great diversity. As in the '30s, writers new and old produced work that was lively and energetic. If they sometimes formed alliances, they also knew the wisdom of D.J. Enright's remark that "the best movement is one that doesn't move far in the same direction." It is difficult to discern any real populism in the writing of the '50s. The revolution, if there was one, was founded on honest impatience with the cant and hypocrisy, genteel pretentiousness, and shifty incompetences that find shelter in the trenches of carelessly perpetuated privilege. The iconoclastic writers of the '50s resembled those of the '30s. They valued common sense, order, clarity, wit, and self-knowledge, and they were not scrupulous in their conduct toward those who appeared to lack these qualities. In later decades, when, as Orwell had ruefully predicted, cant and hypocrisy became as much a weapon of the left as of the right, it is hardly surprising that many of the Angry Young Men were to be found on the opposite side of the barricades.

Novels of Amis, Murdoch, and Wilson

Prose. Kingsley Amis, at different stages of his life the writer of official pamphlets on behalf of both the Labour and Conservative parties, was by far the best of the iconoclastic novelists to emerge in the '50s. (The others included John Wain, John Braine, and Alan Sillitoe.) *Lucky Jim* (1953) established from the first Amis' continuing qualities—his baleful comic stare, his uncanny ear for the quirks of contemporary speech, and the enormous cunning with which he establishes the authority of a central persona who is entirely ordinary (and beset with the usual failings) yet also improbably alert to the most telling distinctions and deceptions in the life lived around him. Like Waugh, Amis is a relentless tease who courts outrage and rejection. His weaker novels lapse sometimes into facetiousness, but the uniqueness of this persona, more interesting than any question of ideology, has been remarkably consistent from *Lucky Jim* to *Stanley and the Women* (1984).

Amis remained a realist in the novel, a contemporary practitioner of the long, liberal humanist tradition, which puts the novelist's imaginative and linguistic gifts primarily at the service of character and therefore of life. His two most important contemporaries, Iris Murdoch and Sir Angus Wilson, also stand solidly within that tradition, but they are more concerned with the structural problems of writing such novels at a time when society no longer supplies a coherent and communal pattern of experience. Iris Murdoch's novels explore good and evil, the nature of power, and the possibility of love from a standpoint that is neither traditional nor modernist but creates its own flavour, as do the novels of Dame Ivy Compton-Burnett. Murdoch's lively gift of social observation is powerfully, sometimes grotesquely, combined with an insistent psychological, sexual, or mythical patterning: it is as though the transparent surface of the comedy of manners has been stretched and darkened to embrace the calamitous events of tragedy. Even more directly than Iris Murdoch, Sir Angus Wilson takes as his subject the crisis of educated middle-class society in England since World War I. Novelists such as C.P. Snow (*Strangers and Brothers,* 1940–70) treated similar themes in an entirely conventional way, as though they were writing 19th-century novels in modern dress. But if Snow's sequence is compared with Wilson's panoramic novel *No Laughing Matter* (1967), with its ingenious combination of realism, fantasy, and parody, it will be seen that it is Wilson who has contrived to tell the reader more about reality.

Such accommodations between the novel of character and the pressures of contemporary fictional theory have been extremely important: they have preserved the English novel's relation to life (elsewhere fiction sometimes risks being swallowed up by an ever more complex body of theory) while renewing its vitality as an art. A welcome feature of the period is the cross-fertilization between the newer writers and a number of older novelists. Ivy Compton-Burnett has already been mentioned. Other examples are Elizabeth Bowen (*The Heat of the Day,* 1949; *A World of Love,* 1955); L.P. Hartley (*The Go-Between,* 1953; *The Hireling,* 1957); Nancy Mitford (*The Blessing,* 1951; *Don't Tell Alfred,* 1960); Rebecca West (*The Fountain Overflows,* 1956; *The Birds Fall Down,* 1966); Jean Rhys (*Wide Sargasso Sea,* 1966); Richard Hughes (*The Human Predicament,* 1961–73); Olivia Manning (*The Balkan Trilogy,* 1960–65); Elizabeth Taylor (*Angel,* 1957); and Sybille Bedford (*A Legacy,* 1956). Realism, arguably, is what the English novel does best, but complacency on that score can lead to a deadly predictability. Yet Iris Murdoch's insistence that "the function of the writer [is] to write the best books he knows how to write" is a salutary reminder of the strength of the intuitive English tradition. In the 1960s and after, the better novelists may have become more overtly concerned with problems of form, but they still derived strength and variety from grafting these concerns onto their firm adherence to the liberal novel, in marked contrast to the arid pretentiousness of so much continental fiction. Examples of this strength may be found in the work of William Golding (*Lord of the Flies,* 1954; *Pincher Martin,* 1956; *Rites of Passage,* 1980); Anthony Burgess (*A Clockwork Orange,* 1962; *Inside Mr. Enderby,* 1963; *Earthly Powers,* 1980); Doris Lessing (*The Golden Notebook,* 1962); Muriel Spark (*Memento Mori,* 1959; *The Driver's Seat,* 1970; *The Takeover,* 1976); and V.S. Naipaul (*A House for Mr. Biswas,* 1961; *The Mimic Men,* 1967; *In a Free State,* 1971).

Poetry. The state of poetry in the 1950s in many ways

resembled that of the novel; indeed the period was remarkable in that Fuller, Larkin, Wain, Amis, and Enright were competent practitioners in both genres. These poets formed the nucleus of a group briefly known as The Movement, represented in Robert Conquest's anthology *New Lines* (1956). As ever, the grouping conceals diversity, but it is largely true that the Movement poets cultivated plainsong rather than polyphony. The typical Movement poem (as A. Alvarez demonstrated with a roguish concoction of disparate lines in his anti-Movement anthology *The New Poetry,* 1962) is wryly sober, reflectively public, and somewhat flat. It parades the poet as ordinary man, his tones and themes defined by his membership in the educated middle class. The poet who partakes of the Movement's virtues but transcends its defects is Philip Larkin. Larkin uses words and rhythms to paint subtle and individual impressions on the mind. He has a rare gift of control and organization, and he can distill beauty from ordinariness and dullness and pain, renewing the force of Keats's assertion that "Beauty is truth, truth beauty." He is the latest in that honourable line of English poets who demonstrate that greatness can speak in a minor key. In the '60s those who found the Movement poets limited and insular championed the work of Ted Hughes (who succeeded Betjeman as poet laureate in 1984) with its controlled violence, inwardness with animal instinct, and powerful investigation of the untamed part of human personality. Also interesting are the more compact, tautly elegant poet Thom Gunn; the Australian poet Peter Porter; and the learned, ambiguous, and resonant poetry of Geoffrey Hill. The poet who built most fruitfully on the Larkin–Hughes dichotomy of the '60s, gradually acquiring his own oblique but authoritative voice, is the Irishman Seamus Heaney, who emerged in the 1970s as the strongest and most influential poet writing today. His works include *North,* 1975; *Field Work,* 1979; and *Station Island,* 1984.

Drama. It was in the theatre, far more than in poetry or the novel, that genuinely radical changes took place in the postwar period. In *Look Back in Anger* (1957) John Osborne was determined to ignore the suffocating conventions of the well-made West End play in order to allow the pressures and excitements of contemporary life to be heard in the theatre. In the process he developed a heightened realism that, in plays such as *The Entertainer* (1957) and *A Patriot for Me* (1966), combined the passionate immediacy of drama with the complex psychological portraiture of the novel. Osborne's revolution was one of scale, attitude, and feeling rather than formal invention, but the revolution of method inaugurated by the Irish playwright Samuel Beckett was even more important. *Waiting for Godot* (1954) showed how the simplest of means (a bare stage, a tree, lighting) and the sparest of verbal exchanges could move and delight audiences by dramatizing the interior, subverbal workings of the human spirit. Beckett refused the seductive illusions of art, seeking instead the simplest and most essential bones of meaning. This minimalist stance has obvious limitations, and Beckett's subsequent plays *Endgame* (1958), *Happy Days* (1961), *Breath* (1969), and *Not I* (1973) lovingly described them, until his ultimate dramatic goal came to be a pregnant silence. Harold Pinter—author of *The Birthday Party,* 1958; *The Caretaker,* 1960; and *The Homecoming,* 1965, among other plays—learned from Beckett the art of writing plays that give nothing away, but he developed it in a highly distinctive manner. Where more conventional dramatists feel compelled to explain what in life is often opaque or incommunicable, Pinter shows that there is another drama to be made out of blankly presenting the psychic power struggles and neurotic menace that everyone experiences without understanding. Something of this oblique and slightly surreal vision survives in the bold and overtly radical drama of Edward Bond (*The Pope's Wedding,* 1962; *Saved,* 1965; *Lear,* 1971), with its contrasting of contemporary horror with the possibility of innocence and redemption. Politics surface again in the rambunctious and anarchical comedies of Joe Orton (*Entertaining Mr. Sloane,* 1963; *Loot,* 1964; and *What the Butler Saw,* 1967). Stylistic invention of a very different kind was to be seen in the plays of Tom Stoppard (*Rosencrantz and Guildenstern Are Dead,* 1967; *Jumpers,* 1972; *Travesties,* 1975), a linguistic dandy whose virtuosity gives sparkling theatrical form to his philosophical concern with questions of art and reality.

An important characteristic of the period following 1975 is its undirected openness to different styles of writing and different perceptions of reality. In the theatre, which has remained the most political genre, several impressive heirs to the radical tradition such as Trevor Griffiths, Howard Brenton, David Hare, and David Edgar have widened their dramatic scope with techniques adapted from films and television, but similar skills have been as richly exploited by bourgeois ironists such as Simon Gray, Alan Bennett, Michael Frayn, Alan Ayckbourn, and Stoppard. In the novel the rapturous reception accorded to such self-consciously avant-garde writers as D.M. Thomas (*The White Hotel,* 1981) and Salman Rushdie (*Midnight's Children,* 1981) suggests an intellectual hunger for experimentation, but at the same time due recognition is given to the careful and utterly unpretentious fiction of Anita Brookner (*A Start in Life,* 1981; *Hotel du Lac,* 1984). It is notable that, although certain important modernist achievements have come to be more widely respected in recent years, their influence has never really taken root. Examples are provided by the novels of Beckett (*Murphy,* 1938; *Molloy,* 1955; *Malone Dies,* 1956), the poetry of Hugh MacDiarmid and Basil Bunting, and the work of David Jones (*In Parenthesis,* 1937; *The Anathemata,* 1952).

In the 1960s it briefly seemed possible that the traditional roots of English writing would be significantly shifted—that literature would become less metropolitan and academic and would answer more closely to the needs of the young, the working class, and the ideologically committed. This trend was particularly evident in poetry (pop poetry, performance poetry, protest poetry), but with hindsight it is clear that its influence was short-lived and limited and that contemporary writing has retreated, if not to the ivory tower, then at least to the "palace of art." It is a symptom of the change that new writers such as the poets Douglas Dunn and Tony Harrison actually emphasized the tension between the practice of their art and their commitment to their working-class origins, while the most interesting new writing exhibits a dandyish delight in verbal and formal manipulations combined with a serious underlying concern with the nature of art and the individual imagination. Examples may be found in the "Martian" poetry of Craig Raine (*The Onion Memory,* 1978; *A Martian Sends a Postcard Home,* 1979; *Rich,* 1984), the novels of Martin Amis (especially *Money,* 1984), and Stoppard's play *The Real Thing* (1982). Writing, it may be concluded, has remained an inescapably bourgeois and liberal activity, but, as such, it faces serious difficulties both within and without. Social and economic changes have progressively turned the retreat of writers toward the universities, which began in the 1950s, into a stampede. Meanwhile, in spite of the publicity and enthusiasm created by literary prizes, the habit of reading is threatened on all sides, by television, by falling standards of education, and by the erosion of communal myths or frames of reference. (F.E.D.)

The Movement

Openness to diversity in the theatre

Radical changes in the theatre

BIBLIOGRAPHY

General works: A comprehensive reference source with emphasis on British authors and their writings is the *Oxford Companion to English Literature.* The 4th edition, edited by SIR PAUL HARVEY and revised by DOROTHY EAGLE (1969), and the 5th edition, edited by MARGARET DRABBLE (1985), have somewhat different but overlapping coverage. *The Oxford History of English Literature,* a multivolume series, provides comprehensive coverage of the subject, as does *The Cambridge History of English Literature,* edited by A.W. WARD and A.R. WALLER, 15 vol. (1907–33, reprinted 1976). Another useful source is PETER CONRAD, *The Everyman History of English Literature* (1985).

Old English and Early Middle English literature: DEREK PEARSALL, *Old English and Middle English Poetry* (1977), is a good critical survey of both periods. STANLEY B. GREENFIELD and FRED C. ROBINSON, *A Bibliography of Publications on Old English Literature to the End of 1972* (1980, reprinted 1982), lists more than 6,500 items covering published materials on the subject and has a good index. STANLEY B. GREENFIELD and DANIEL G. CALDER, *A New Critical History of Old English Literature* (1986), serves as a good introductory survey. GEORGE

PHILIP KRAPP and ELLIOTT VAN KIRK DOBBIE (eds.), *The Anglo-Saxon Poetic Records,* 6 vol. (1931–53), is the standard edition of Old English poetry; and S.A.J. BRADLEY (ed. and trans.), *Anglo-Saxon Poetry* (1982), is an anthology of Old English poems in prose translation. R.M. WILSON, *Early Middle English Literature,* 3rd ed. (1968), is a critical survey of the period; J. BURKE SEVERS and ALBERT E. HARTUNG (eds.), *A Manual of the Writings in Middle English, 1050–1500* (1967–), a multivolume reference, seven volumes of which were published by 1986, contains commentaries on the individual works and extensive bibliographies; and J.A.W. BENNETT and G.V. SMITHERS (eds.), *Early Middle English Verse and Prose,* 2nd ed. (1968, reprinted 1982), is an authoritative anthology, with a glossary.

The Middle English period: A.S.G. EDWARDS (ed.), *Middle English Prose: A Critical Guide to Major Authors and Genres* (1984), includes bibliographies and surveys of scholarship. See also A.S.G. EDWARDS and DEREK PEARSALL (eds.), *Middle English Prose: Essays on Bibliographical Problems* (1981). BORIS FORD (ed.), *The New Pelican Guide to English Literature,* vol. 1, *Medieval Literature,* Part One: *Chaucer and the Alliterative Tradition,* rev. ed. (1982), is an anthology with extensive bibliographies; another valuable source is CARL J. STRATMAN, *Bibliography of Medieval Drama,* 2nd ed. rev. and enlarged, 2 vol. (1972). Analytical studies include PIERO BOITANI, *English Medieval Narrative in the Thirteenth and Fourteenth Centuries* (1982; originally published in Italian, 1980); J.A. BURROW, *Ricardian Poetry: Chaucer, Gower, Langland, and the Gawain Poet* (1971); PAMELA GRADON, *Form and Style in Early English Literature* (1971, reprinted 1974); RICHARD FIRTH GREEN, *Poets and Princepleasers: Literature and the English Court in the Late Middle Ages* (1980); DAVID LAWTON (ed.), *Middle English Alliterative Poetry and Its Literary Background* (1982); CHARLES MUSCATINE, *Poetry and Crisis in the Age of Chaucer* (1972); ROBERT A. POTTER, *The English Morality Play: Origins, History, and Influence of a Dramatic Tradition* (1975); V.J. SCATTERGOOD, *Politics and Poetry in the Fifteenth Century* (1971); A.C. SPEARING, *Medieval Dream-Poetry* (1976), and *Medieval to Renaissance in English Poetry* (1985); R.M. WILSON, *Lost Literature of Medieval England,* 2nd rev. ed. (1970); GEORGE KANE, *Middle English Literature* (1951, reissued 1979); DOROTHY EVERETT, *Essays on Middle English Literature* (1955, reprinted 1978); C.S. LEWIS, *The Allegory of Love: A Study in Medieval Tradition* (1936, reprinted 1977); DIETER MEHL, *The Middle English Romances of the Thirteenth and Fourteenth Centuries* (1969; originally published in German, 1967); ARTHUR K. MOORE, *The Secular Lyric in Middle English* (1951, reprinted 1970); ROSEMARY WOOLF, *The English Religious Lyric in the Middle Ages* (1968); ROBERTO WEISS, *Humanism in England During the Fifteenth Century,* 3rd ed. (1967); M.J.C. HODGART, *The Ballads,* 2nd ed. (1964); THORLAC TURVILLE-PETRE, *The Alliterative Revival* (1977); and ROSEMARY WOOLF, *The English Mystery Plays* (1972, reissued 1980).

The Renaissance and the 17th century: (*Elizabethan poetry and prose*): C.S. LEWIS, *English Literature in the Sixteenth Century: Excluding Drama* (1954, reprinted 1973), is a standard literary survey. CHRISTOPHER RICKS (ed.), *English Poetry and Prose, 1540–1674* (1970); and BORIS FORD (ed.), *The Age of Shakespeare,* rev. ed. (1982), are useful collections of essays. FRANCES A. YATES, *Astraea: The Imperial Theme in the Sixteenth Century* (1975); JOHN BUXTON, *Elizabethan Taste* (1963); and LOUIS B. WRIGHT, *Middle-Class Culture in Elizabethan England* (1935, reissued 1980), explore the backgrounds of literature, including politics and patronage of the period. Specific topics are discussed in HALLETT SMITH, *Elizabethan Poetry: A Study in Conventions, Meaning, and Expression* (1952, reprinted 1964); PAUL J. ALPERS (ed.), *Elizabethan Poetry: Modern Essays in Criticism* (1967); FRANK KERMODE, *Shakespeare, Spenser, Donne: Renaissance Essays* (1971; U.K. title, *Renaissance Essays: Shakespeare, Spenser, Donne,* 1973); DOUGLAS L. PETERSON, *The English Lyric from Wyatt to Donne* (1967); J.W. LEVER, *The Elizabethan Love Sonnet,* 2nd ed. (1966); and ROSEMOND TUVE, *Elizabethan and Metaphysical Imagery* (1947, reissued 1968), on rhetoric. Anthologies include NORMAN AULT (ed.), *Elizabethan Lyrics from the Original Texts,* 4th ed. (1966); NIGEL ALEXANDER (ed.), *Elizabethan Narrative Verse* (1967); and EDWARD LUCIE-SMITH (ed.), *The Penguin Book of Elizabethan Verse* (1965). (*Elizabethan and early Stuart drama*): The theatrical background is surveyed in ANDREW GURR, *The Shakespearean Stage, 1574–1642,* 2nd ed. (1980); and KENNETH MUIR and S. SCHOENBAUM (eds.), *A New Companion to Shakespeare Studies* (1971, reprinted 1976). CHRISTOPHER RICKS (ed.), *English Drama to 1710* (1971), is a collection of essays. Surveys of the literature include MADELEINE DORAN, *Endeavors of Art: A Study of Form in Elizabethan Drama* (1954, reprinted 1964); M.C. BRADBROOK, *The Growth and Structure of Elizabethan Comedy,* new ed. (1973, reprinted 1979); and ALFRED HARBAGE, *Shakespeare and the Rival Traditions* (1952, reissued 1968), on the companies of schoolboy actors. The following are special studies: J.M.R. MARGESON, *The Origins of English Tragedy* (1967); DAVID BEVINGTON, *From Mankind to Marlowe: Growth of Structure in the Popular Drama of Tudor England* (1962); C.L. BARBER, *Shakespeare's Festive Comedy: A Study of Dramatic Form and Its Relation to Social Custom* (1959, reprinted 1972); E.M.W. TILLYARD, *Shakespeare's History Plays* (1944, reprinted 1980); A.C. BRADLEY, *Shakespearean Tragedy: Lectures on Hamlet, Othello, King Lear, Macbeth,* 2nd ed. (1985); BRIAN GIBBONS, *Jacobean City Comedy,* 2nd ed. (1980); MARGOT HEINEMANN, *Puritanism and Theatre: Thomas Middleton and Opposition Drama Under the Early Stuarts* (1980, reprinted 1982); J.W. LEVER, *The Tragedy of State: A Study in Jacobean Drama* (1971, reprinted 1980); and ENID WELSFORD, *The Court Masque: A Study in the Relationship Between Poetry and Revels* (1927, reissued 1962). (*Early Stuart poetry and prose*): DOUGLAS BUSH, *English Literature in the Earlier Seventeenth Century,* 2nd rev. ed. (1962, reprinted 1976), is a standard survey. Historical background is explored in HIRAM HAYDN, *The Counter-Renaissance* (1950, reprinted 1966); ALAN SINFIELD, *Literature in Protestant England, 1560–1660* (1983); CHRISTOPHER HILL, *Intellectual Origins of the English Revolution* (1965, reprinted with corrections 1980), and *The World Turned Upside Down: Radical Ideas During the English Revolution* (1972, reissued 1982); and BASIL WILLEY, *The Seventeenth Century Background: Studies in the Thought of the Age in Relation to Poetry and Religion* (1934, reprinted 1977). Information on the court is found in D.J. GORDON, *The Renaissance Imagination* (1975); and ROY STRONG, *Van Dyck: Charles I on Horseback* (1972). Special topical studies include LOUIS L. MARTZ, *The Poetry of Meditation: A Study in English Religious Literature of the Seventeenth Century,* rev. ed. (1962, reprinted 1978); MAREN-SOFIE RØSTVIG, *The Happy Man: Studies in the Metamorphoses of a Classical Ideal,* 2nd ed., 2 vol. (1962–71), on the Cavalier poetry; C.A. PATRIDES and RAYMOND B. WADDINGTON (eds.), *The Age of Milton: Backgrounds to Seventeenth-Century Literature* (1980); BRIAN VICKERS, *Essential Articles for the Study of Francis Bacon* (1968); JOAN WEBBER, *The Eloquent "I": Style and Self in Seventeenth-Century Prose* (1968); and STANLEY E. FISH (ed.), *Seventeenth-Century Prose: Modern Essays in Criticism* (1971).

The Restoration and the 18th century: Helpful introductions to the period can be found in the relevant volumes of the Oxford History of English Literature series: JAMES SUTHERLAND, *English Literature of the Late Seventeenth Century* (1969); BONAMY DOBRÉE, *English Literature in the Early Eighteenth Century, 1700–1740* (1959, reprinted with corrections 1964); and JOHN BUTT, *The Mid-Eighteenth Century,* ed. and completed by GEOFFREY CARNALL (1979); and in such monographs as A.R. HUMPHREYS, *The Augustan World: Life and Letters in Eighteenth-Century England* (1954, reprinted 1978); MAXIMILLIAN E. NOVAK, *Eighteenth-Century English Literature* (1983); PAT ROGERS, *The Augustan Vision* (1974, reissued 1978); PAT ROGERS (ed.), *The Eighteenth Century* (1978); LESLIE STEPHEN, *English Literature and Society in the Eighteenth Century* (1904, reprinted 1965); and STEPHEN COPLEY (ed.), *Literature and the Social Order in Eighteenth-Century England* (1984). Useful studies that focus on more restricted topics but cover the whole of the period include HOWARD ERSKINE-HILL, *The Augustan Idea in English Literature* (1983); PAUL FUSSELL, *The Rhetorical World of Augustan Humanism: Ethics and Imagery from Swift to Burke* (1965, reprinted 1969); JEAN H. HAGSTRUM, *Sex and Sensibility: Ideal and Erotic Love from Milton to Mozart* (1980); IAN JACK, *Augustan Satire: Intention and Idiom in English Poetry, 1660–1750* (1952, reprinted 1966); JAMES WILLIAM JOHNSON, *The Formation of English Neo-Classical Thought* (1967, reprinted 1978); MARTIN PRICE, *To the Palace of Wisdom: Studies in Order and Energy from Dryden to Blake* (1964, reissued 1970); ERIC ROTHSTEIN, *Restoration and Eighteenth-Century Poetry, 1660–1780* (1981); JAMES SUTHERLAND, *A Preface to Eighteenth-Century Poetry* (1948, reprinted 1970); and RACHEL TRICKETT, *The Honest Muse: A Study in Augustan Verse* (1967). Among important thematic or general studies with a narrower chronological range are JOHN BARRELL, *English Literature in History, 1730–80* (1983); WALTER JACKSON BATE, *From Classic to Romantic: Premises of Taste in Eighteenth-Century England* (1946, reprinted 1961); DONALD DAVIE, *A Gathered Church: The Literature of the English Dissenting Interest, 1700–1930* (1978); CHRISTOPHER HILL, *The Experience of Defeat: Milton and Some Contemporaries* (1984); EARL MINER, *The Restoration Mode from Milton to Dryden* (1974); SAMUEL H. MONK, *The Sublime: A Study of Critical Theories in XVIII-Century England* (1935, reissued 1960); MARJORIE HOPE NICOLSON, *Newton Demands the Muse: Newton's Opticks and the Eighteenth Century Poets* (1946, reprinted 1979), and *Science and Imagination* (1956, reprinted 1976); RONALD PAULSON, *Satire and the Novel in Eighteenth-Century England* (1967); JOHN PRESTON, *The Created Self: The Reader's Role in Eighteenth-Century Fiction* (1970); JOHN J. RICHETTI,

Popular Fiction Before Richardson: Narrative Patterns, 1700–1739 (1969); PAT ROGERS, *Hacks and Dunces: Pope, Swift, and Grub Street* (1980); PATRICIA MEYER SPACKS, *Imagining a Self: Autobiography and Novel in Eighteenth-Century England* (1976), and *The Poetry of Vision: Five Eighteenth-Century Poets* (1967); GEOFFREY TILLOTSON, *Augustan Poetic Diction* (1964); IAN WATT, *The Rise of the Novel: Studies in Defoe, Richardson, and Fielding* (1957, reprinted 1971); BASIL WILLEY, *The Eighteenth Century Background: Studies on the Idea of Nature in the Thought of the Period* (1940, reprinted 1980); and JOHN HAROLD WILSON, *The Court Wits of the Restoration: An Introduction* (1948, reissued 1967). Interesting explorations of individual major writers include MAXIMILLIAN E. NOVAK, *Defoe and the Nature of Man* (1963); G.A. STARR, *Defoe and Spiritual Autobiography* (1965, reissued 1971); C.J. RAWSON, *Henry Fielding and the Augustan Ideal Under Stress: "Nature's Dance of Death" and Other Studies* (1972); WALTER JACKSON BATE, *Samuel Johnson* (1977); REUBEN A. BROWER, *Alexander Pope: The Poetry of Allusion* (1959, reprinted 1968); MAYNARD MACK, *The Garden and the City: Retirement and Politics in the Later Poetry of Pope, 1731–1743* (1969), and *Alexander Pope: A Life* (1985); AUBREY L. WILLIAMS, *Pope's "Dunciad": A Study of Its Meaning* (1955, reprinted 1968); MARGARET ANNE DOODY, *A Natural Passion: A Study of the Novels of Samuel Richardson* (1974); DAVID M. VIETH, *Attribution in Restoration Poetry: A Study of Rochester's Poems of 1680* (1963); and IRVIN EHRENPREIS, *Swift: The Man, His Works, and the Age,* 3 vol. (1962–83). Theatrical history is chronicled in ROBERT D. HUME, *The Development of English Drama in the Late Seventeenth Century* (1976); PETER HOLLAND, *The Ornament of Action: Text and Performance in Restoration Comedy* (1979); RICHARD BEVIS, *The Laughing Tradition: Stage Comedy in Garrick's Day* (1980); and ARTHUR SHERBO, *English Sentimental Drama* (1957). Among collections and anthologies are DONALD DAVIE (ed.), *Augustan Lyric* (1974); H.T. DICKINSON (ed.), *Politics and Literature in the Eighteenth Century* (1974); SCOTT ELLEDGE (ed.), *Eighteenth-Century Critical Essays,* 2 vol. (1961); H.G.C. GRIERSON and G. BULLOUGH (eds.), *Oxford Book of Seventeenth Century Verse* (1934, reprinted 1976); DAVID W. LINDSAY (ed.), *English Poetry, 1700–1780: Contemporaries of Swift and Johnson* (1974); GEORGE DE F. LORD *et al.* (eds.), *Poems on Affairs of State: Augustan Satirical Verse, 1660–1714,* 7 vol. (1963–75); HAROLD LOVE (ed.), *The Penguin Book of Restoration Verse* (1968); GEORGE H. NETTLETON and ARTHUR E. CASE (eds.), *British Dramatists from Dryden to Sheridan* (1939, reprinted 1975); DAVID NICHOL SMITH (ed.), *Characters from the Histories and Memoirs of the Seventeenth Century* (1918, reprinted 1967), *Eighteenth Century Essays on Shakespeare,* 2nd ed. (1963), and *The Oxford Book of Eighteenth Century Verse* (1926, reprinted 1965); ROGER LONSDALE (ed.), *The New Oxford Book of Eighteenth Century Verse* (1984); CHARLES PEAKE (ed.), *Poetry of the Landscape and the Night: Two Eighteenth-Century Traditions* (1967, reissued 1970); FRANCIS VENABLES (ed.), *The Early Augustans* (1972); and TIMOTHY WEBB (ed.), *English Romantic Hellenism, 1700–1824* (1982).

The Romantic period: General literary history of the period is presented in the relevant volumes of the Oxford History of English Literature series: W.L. RENWICK, *English Literature, 1789–1815* (1963), and IAN JACK, *English Literature, 1815–1832* (1963); as well as in R.A. FOAKES, *The Romantic Assertion: A Study in the Language of Nineteenth Century Poetry* (1958, reprinted 1971); JOHN O. HAYDEN (ed. and comp.), *Romantic Bards and British Reviewers: A Selected Edition of the Contemporary Reviews of the Works of Wordsworth, Coleridge, Byron, Keats, and Shelley* (1971, reprinted 1976); and THEODORE REDPATH (comp.), *The Young Romantics and Critical Opinion, 1807–1824: Poetry of Byron, Shelley, and Keats as Seen by Their Contemporary Critics* (1973). Social and intellectual background of the period is the subject of numerous works: RAYMOND WILLIAMS, *Culture and Society, 1780–1950* (1958, reprinted 1983); M.H. ABRAMS, *The Mirror and the Lamp: Romantic Theory and the Critical Tradition* (1953, reprinted 1971), and *Natural Supernaturalism: Tradition and Revolution in Romantic Literature* (1971); MARILYN BUTLER, *Jane Austen and the War of Ideas* (1975), and *Romantics, Rebels, and Reactionaries: English Literature and Its Background, 1760–1830* (1981); HERBERT W. PIPER, *The Active Universe* (1962); CARL WOODRING, *Politics in English Romantic Poetry* (1970); H.G. SCHENK, *The Mind of the European Romantics: An Essay in Cultural History* (1966, reissued 1969); STEPHEN PRICKETT, *The Romantics* (1981), and *Romanticism and Religion: The Tradition of Coleridge and Wordsworth in the Victorian Church* (1976); and LILIAN R. FURST, *Romanticism in Perspective: A Comparative Study of Aspects of the Romantic Movements in England, France, and Germany,* 2nd ed. (1979). Analytical studies of narrower topics include J.R. WATSON, *English Poetry of the Romantic Period, 1789–1830* (1985), and *Picturesque Landscape and English Romantic Poetry* (1970); THOMAS WEISKEL, *The Romantic Sublime: Studies in the Structure and Psychology of Transcendence* (1976); THOMAS MCFARLAND, *Romanticism and the Forms of Ruin: Wordsworth, Coleridge, and Modalities of Fragmentation* (1981), and *Coleridge and the Pantheist Tradition* (1969); ELIZABETH SEWELL, *The Orphic Voice: Poetry and Natural History* (1960, reissued 1971); C.M. BOWRA, *The Romantic Imagination* (1949, reissued 1969); MICHAEL G. COOKE, *The Romantic Will* (1976); HAROLD BLOOM, *The Visionary Company: A Reading of English Romantic Poetry,* rev. ed. (1971), and *Poetry and Repression: Revisionism from Blake to Stevens* (1976); PAUL A. CANTOR, *Creature and Creator: Myth-Making and English Romanticism* (1984); G. WILSON KNIGHT, *The Starlit Dome: Studies in the Poetry of Vision,* rev. ed. (1959, reissued 1971); DAVID MORSE, *Perspectives on Romanticism: Transformational Analysis* (1981), and *Romanticism, a Structural Analysis* (1982); ALBERT S. GÉRARD, *English Romantic Poetry: Ethos, Structure, and Symbol in Coleridge, Wordsworth, Shelley, and Keats* (1968); and DAVID G. JAMES, *The Romantic Comedy* (1948, reprinted 1980). Comprehensive collections are represented by H.S. MILFORD (comp.), *The Oxford Book of English Verse of the Romantic Period, 1798–1830* (1935, reprinted 1974); HAROLD BLOOM (ed.), *English Romantic Poetry: An Anthology,* 2 vol. (1963); and HAROLD BLOOM and LIONEL TRILLING (eds.), *Romantic Poetry and Prose* (1973).

Victorian literature: Comprehensive studies of the period, introducing the literary background, include WALTER E. HOUGHTON, *The Victorian Frame of Mind, 1830–1870* (1957, reprinted 1966); JEROME HAMILTON BUCKLEY, *The Victorian Temper: A Study in Literary Culture* (1951, reprinted 1981); BASIL WILLEY, *Nineteenth Century Studies: Coleridge to Matthew Arnold* (1949, reissued 1980); ARTHUR POLLARD (ed.), *The Victorians* (1970); and G.K. CHESTERTON, *The Victorian Age in Literature* (1913). Studies of special subjects are presented in GEORGE LEVINE and WILLIAM MADDEN (eds.), *The Art of Victorian Prose* (1968), on nonfiction; MICHAEL WHEELER, *English Fiction of the Victorian Period: 1830–1890* (1985); KATHLEEN TILLOTSON, *Novels of the Eighteen-Forties* (1954, reprinted 1983); ROBERT LANGBAUM, *The Poetry of Experience: The Dramatic Monologue in Modern Literary Tradition* (1957, reprinted 1985), on Victorian poetry; GEORGE ROWELL, *The Victorian Theatre, 1792–1914,* 2nd ed. (1978); and ROGER B. HENKLE, *Comedy and Culture, 1820–1900* (1980).

Modern English literature (1900–45): MALCOLM BRADBURY, *The Social Context of Modern English Literature* (1971), is a discussion of the effects of modernization on the form and content of English literature and on the role of the modern writer. MALCOLM BRADBURY and JAMES MCFARLANE (eds.), *Modernism: 1890–1930* (1976), is a collection of essays focusing on the international context of Anglo-American modernism. MICHAEL H. LEVENSON, *A Genealogy of Modernism: A Study of English Literary Doctrine, 1908–1922* (1984), is a meticulously detailed history of the modernist movement in England. *From James to Eliot* (1983) is a comprehensive collection of essays on the social and cultural background of the literature of the early 20th century; it is the seventh volume in BORIS FORD (ed.), *The New Pelican Guide to English Literature,* 8 vol., rev. ed. (1982–83). CHRISTOPHER GILLIE, *Movements in English Literature, 1900–1940* (1975), is a straightforward introduction to the history of fiction, poetry, and drama of the period. The historical background is also explored in SAMUEL HYNES, *The Auden Generation: Literature and Politics in England in the 1930s* (1976, reissued 1982); and ROBERT HEWISON, *Under Siege: Literary Life in London 1939–1945* (1977). DAVID PERKINS, *A History of Modern Poetry: From the 1890s to the High Modernist Mode* (1976), is a broad study with emphasis on the interplay between British and American poetry. JOHN PRESS, *A Map of Modern English Verse* (1969), is an analysis of traditional and modernist poetry from the 1900s to the 1950s.

Literature after 1945: ROBERT HEWISON, *In Anger: Culture in the Cold War, 1945–60* (1981), is a general historical study of the period; and BORIS FORD (ed.), *The Present* (1983), is a sound general survey. BERNARD BERGONZI, *The Situation of the Novel,* 2nd ed. (1979), presents a comprehensive treatment of contemporary fiction; MALCOLM BRADBURY and DAVID PALMER (eds.), *The Contemporary English Novel* (1980), reflects the critical opinion of the 1970s; as does MALCOLM BRADBURY (ed.), *The Novel Today: Contemporary Writers on Modern Fiction* (1977). JOHN ELSOM, *Post-War British Theatre,* rev. ed. (1979), is a historical survey; and JOHN RUSSELL BROWN, *A Short Guide to Modern British Drama* (1983), gives brief accounts of general trends and individual plays. EDWARD LUCIE-SMITH (ed.), *British Poetry Since 1945* (1970); D.J. ENRIGHT (ed.), *The Oxford Book of Contemporary Verse, 1945–1980* (1980); and BLAKE MORRISON and ANDREW MOTION (eds.), *The Penguin Book of Contemporary British Poetry* (1982), are representative anthologies with critical information on the movements and trends.

(P.S.Ba./Ri.B./M.H.B./M.Co./J.B.B./N.Sh./H.A.Da./F.E.D.)

Epistemology

Epistemology is one of the main branches of philosophy; its subject matter concerns the nature, origin, scope, and limits of human knowledge. The name is derived from the Greek terms *epistēmē* (knowledge) and *logos* (theory), and accordingly this branch of philosophy is also referred to as the theory of knowledge.

For coverage of related topics in the *Macropædia* and the *Micropædia,* see the *Propædia,* sections 10/51, 10/52, and 10/53, and the *Index.*

This article is divided into the following sections:

Issues of epistemology

EPISTEMOLOGY AS A DISCIPLINE

Why should there be such a subject as epistemology? Aristotle provided the answer when he said that philosophy begins in wonder, in a kind of puzzlement about things. Nearly all human beings wish to comprehend the world they live in, a world that includes the individual as well as other persons, and most people construct hypotheses of varying degrees of sophistication to help them make sense of that world. No conjectures would be necessary if the world were simple; but its features and events defy easy explanation. The ordinary person is likely to give up somewhere in the process of trying to develop a coherent account of things and to rest content with whatever degree of understanding he has managed to achieve.

Philosophers, in contrast, are struck by, even obsessed by, matters that are not immediately comprehensible. Philosophers are, of course, ordinary persons in all respects except perhaps one. They aim to construct theories about the world and its inhabitants that are consistent, synoptic, true to the facts and that possess explanatory power. They thus carry the process of inquiry further than people generally tend to do, and this is what is meant by saying that they have developed a philosophy about these matters. Epistemologists, in particular, are philosophers whose theories deal with puzzles about the nature, scope, and limits of human knowledge.

Like ordinary persons, epistemologists usually start from the assumption that they have plenty of knowledge about the world and its multifarious features. Yet, as they reflect upon what is presumably known, epistemologists begin to discover that commonly accepted convictions are less secure than originally assumed and that many of man's firmest beliefs are dubious or possibly even chimerical. Such doubts and hesitations are caused by anomalous features of the world that most people notice but tend to minimize or ignore. Epistemologists notice these things too, but, in wondering about them, they come to realize that they provide profound challenges to the knowledge claims that most individuals blithely and unreflectingly accept as true.

What then are these puzzling issues? While there is a vast array of such anomalies and perplexities, which will be discussed below in the section on the history of epistemology, two of these issues will be briefly described in order to illustrate why such difficulties call into question common claims to have knowledge about the world.

TWO EPISTEMOLOGICAL PROBLEMS

"Our knowledge of the external world." Most people have noticed that vision can play tricks on them. A straight stick put in water looks bent to them, but they know it is not; railroad tracks are seen to be converging in the distance, yet one knows that they are not; the wheels of wagons on a movie screen appear to be going backward, but one knows that they are not; and the pages of English-language books reflected in mirrors cannot be read from left to right, yet one knows that they were printed to be read that way. Each of these phenomena is thus misleading in some way. If human beings were to accept the world as being exactly as it looks, they would be mistaken about how things really are. They would think the stick in water really to be bent, the railway tracks really to be convergent, and the writing on pages really to be reversed.

Visual anomalies

These are visual anomalies, and they produce the sorts of epistemological disquietudes referred to above. Though they may seem to the ordinary person to be simple problems, not worth serious notice, for those who ponder them they pose difficult questions. For instance, human beings claim to know that the stick is not really bent and the tracks not really convergent. But how do they know that these things are so?

Suppose one says that this is known because, when the stick is removed from the water, one can see that it is not bent. But does seeing a straight stick out of water provide a good reason for thinking that it is not bent when seen in water? How does one know that, when the stick is put into the water, it does not bend? Suppose one says that the tracks do not really converge because the train passes over them at that point. How does one know that the wheels on the train do not happen to converge at that point? What justifies opposing some beliefs to others, especially when

all of them are based upon what is seen? One sees that the stick in water is bent and also that the stick out of the water is not bent. Why is the stick declared really to be straight; why in effect is priority given to one perception over another?

One possible response to these queries is that vision is not sufficient to give knowledge of how things are. One needs to correct vision in some other way in order to arrive at the judgment that the stick is really straight and not bent. Suppose a person asserts that his reason for believing the stick in water is not bent is that he can feel it with his hands to be straight when it is in the water. Feeling or touching is a mode of sense perception, although different from vision. What, however, justifies accepting one mode of perception as more accurate than another? After all, there are good reasons for believing that the tactile sense gives rise to misperception in just the way that vision does. If a person chills one hand and warms the other, for example, and inserts both into a tub of water having a uniform medium temperature, the same water will feel warm to the cold hand and cold to the warm hand. Thus, the tactile sense cannot be trusted either and surely cannot by itself be counted on to resolve these difficulties.

Another possible response is that no mode of perception is sufficient to guarantee that one can discover how things are. Thus, it might be affirmed that one needs to correct all modes of perception by some other form of awareness in order to arrive at the judgment, say, that the stick is really straight. Perhaps that other way is the use of reason. But why should reason be accepted as infallible? It also suffers from various liabilities, such as forgetting, misestimating, or jumping to conclusions. And why should one trust reason if its conclusions run counter to those gained through perception, since it is obvious that much of what is known about the world derives from perception?

Reason as a corrective perception

Clearly there is a network of difficulties here, and one will have to think hard in order to arrive at a clear and defensible explanation of the apparently simple claim that the stick is really straight. A person who accepts the challenge will, in effect, be developing a theory for grappling with the famous problem called "our knowledge of the external world." That problem turns on two issues, namely, whether there is a reality that exists independently of the individual's perception of it—in other words, if the evidence one has for the existence of anything is what one perceives, how can one know that anything exists unperceived?—and, second, how one can know what anything is really like, if the perceptual evidence one has is conflicting.

The "other-minds problem." The second problem also involves seeing but in a somewhat unusual way. It deals with that which one cannot see, namely the mind of another. Suppose a woman is scheduled to have an operation on her right knee and her surgeon tells her that when she wakes up she will feel a sharp pain in her knee. When she wakes up, she does feel the pain the surgeon alluded to. He can hear her groaning and see certain contortions on her face. But he cannot feel what she is feeling. There is thus a sense in which he cannot know what she knows. What he claims to know, he knows because of what others who have undergone operations tell him they have experienced. But, unless he has had a similar operation, he cannot know what it is that she feels.

Indeed, the situation is still more complicated; for, even if the doctor has had such a surgical intervention, he cannot know that what he is feeling after his operation is exactly the same sensation that the woman is feeling. Because each person's sensation is private, the surgeon cannot really know that what the woman is describing as a pain and what he is describing as a pain are really the same thing. For all he knows, she could be referring to a sensation that is wholly different from the one to which he is alluding.

In short, though another person can perceive the physical manifestations the woman exhibits, such as facial grimaces and various sorts of behaviour, it seems that only she can have knowledge of the contents of her mind. If this assessment of the situation is correct, it follows that it is impossible for one person to know what is going on in another person's mind. One can conjecture that a person is experiencing a certain sensation, but one cannot, in a strict sense of the term, know it to be the case.

If this analysis is correct, one can conclude that each human being is inevitably and even in principle cut off from having knowledge of the mind of another. Most people, conditioned by the great advances of modern technology, believe that in principle there is nothing in the world of fact about which science cannot obtain knowledge. But the "other-minds problem" suggests the contrary—namely, that there is a whole domain of private human experience that is resistant to any sort of external inquiry. Thus, one is faced with a profound puzzle, one of whose implications is that there can never be a science of the human mind.

Implications. These two problems resemble each other in certain ways and differ in others, but both have important implications for epistemology.

First, as the divergent perceptions about the stick indicate, things cannot just be as they appear to be. People believe that the stick which looks bent when it is in the water is really straight, and they also believe that the stick which looks straight when it is out of the water is really straight. But, if the belief that the stick in water is really straight is correct, then it follows that the perception human beings have when they see the stick in water cannot be correct. That particular perception is misleading with respect to the real shape of the stick. Hence, one has to conclude that things are not always as they appear to be.

Distinction between appearance and reality

It is possible to derive a similar conclusion with respect to the mind of another. A person can exhibit all the signs of being in pain, but he may not be. He may be pretending. On the basis of what can be observed, it cannot be known with certitude that he is or that he is not in pain. The way he appears to be may be misleading with respect to the way he actually is. Once again vision can be misleading.

Both problems thus force one to distinguish between the way things appear and the way they really are. This is the famous philosophical distinction between appearance and reality. But, once that distinction is drawn, profound difficulties arise about how to distinguish reality from mere appearance. As will be shown, innumerable theories have been presented by philosophers attempting to answer this question since time immemorial.

Second, there is the question of what is meant by "knowledge." People claim to know that the stick is really straight even when it is half-submerged in water. But, as indicated earlier, if this claim is correct, then knowledge cannot simply be identical with perception. For whatever theory about the nature of knowledge one develops, the theory cannot have as a consequence that knowing something to be the case can sometimes be mistaken or misleading.

Third, even if knowledge is not simply to be identified with perception, there nevertheless must be some important relationship between knowledge and perception. After all, how could one know that the stick is really straight unless under some conditions it looked straight? And sometimes a person who is in pain exhibits that pain by his behaviour; thus there are conditions that genuinely involve the behaviour of pain. But what are those conditions? It seems evident that the knowledge that a stick is straight or that one is in great pain must come from what is seen in certain circumstances: perception must somehow be a fundamental element in the knowledge human beings have. It is evident that one needs a theory to explain what the relationship is—and a theory of this sort, as the history of the subject all too well indicates, is extraordinarily difficult to develop.

The two problems also differ in certain respects. The problem of man's knowledge of the external world raises a unique difficulty that some of the best philosophical minds of the 20th century (among them, Bertrand Russell, H.H. Price, C.D. Broad, and G.E. Moore) spent their careers trying to solve. The perplexity arises with respect to the status of the entity one sees when one sees a bent stick in water. In such a case, there exists an entity—a bent stick in water—that one perceives and that appears to be exactly where the genuinely straight stick is. But clearly it cannot be; for the entity that exists exactly where the

straight stick is is the stick itself, an entity that is not bent. Thus, the question arises as to what kind of a thing this bent-stick-in-water is and where it exists.

The responses to these questions have been innumerable, and nearly all of them raise further difficulties. Some theorists have denied that what one sees in such a case is an existent entity at all but have found it difficult to explain why one seems to see such an entity. Still others have suggested that the image seen in such a case is in one's mind and not really in space. But then what is it for something to be in one's mind, where in the mind is it, and why, if it is in the mind, does it appear to be "out there," in space where the stick is? And above all, how does one decide these questions? The various questions posed above only suggest the vast network of difficulties, and in order to straighten out its tangles it becomes indispensable to develop theories.

RELATION OF EPISTEMOLOGY TO OTHER BRANCHES OF PHILOSOPHY

Philosophy viewed in the broadest possible terms divides into many branches: metaphysics, ethics, aesthetics, logic, philosophy of language, philosophy of mind, philosophy of science, and a gamut of others. Each of these disciplines has its special subject matter: for metaphysics it is the ultimate nature of the world; for ethics, the nature of the good life and how people ideally ought to comport themselves in their relations with others; and for philosophy of science, the methodology and results of scientific activity. Each of these disciplines attempts to arrive at a systematic understanding of the issues that arise in its particular domain. The word systematic is important in this connection, referring, as explained earlier, to the construction of sets of principles or theories that are broad-ranging, consistent, and rationally defensible. In effect, such theories can be regarded as sets of complex claims about the various matters that are under consideration.

Epistemology stands in a close and special relationship to each of these disciplines. Though the various divisions of philosophy differ in their subject matter and often in the approaches taken by philosophers to their characteristic questions, they have one feature in common: the desire to arrive at the truth about that with which they are concerned—say, about the fundamental ingredients of the world or about the nature of the good life for man. If no such claims were asserted, there would be no need for epistemology. But, once theses have been advanced, positions staked out, and theories proposed, the characteristic questions of epistemology inexorably follow. How can one know that any such claim is true? What is the evidence in favour of (or against) it? Can the claim be proven? Virtually all of the branches of philosophy thus give rise to epistemological ponderings.

These ponderings may be described as first-order queries. They in turn inevitably generate others that are, as it were, second-order queries, and which are equally or more troubling. What is it to know something? What counts as evidence for or against a particular theory? What is meant by a proof? Or even, as the Greek Skeptics asked, is human knowledge possible at all, or is human access to the world such that no knowledge and no certitude about it is possible? The answers to these second-order questions also require the construction of theories, and in this respect epistemology is no different from the other branches of philosophy. One can thus define or characterize epistemology as that branch of philosophy which is dedicated to the resolution of such first- and second-order queries.

THE NATURE OF KNOWLEDGE

As indicated above, one of the basic questions of epistemology concerns the nature of knowledge. Philosophers normally interpret this query as a conceptual question, *i.e.,* as an issue about a certain conception or idea or notion called knowledge. The question raises a perplexing methodological issue, namely, how does one go about investigating such conceptual questions? It is frequently assumed, though the matter is controversial, that one can determine what knowledge is if one can understand what the word "knowledge" means, that is, what notion or concept the word "knowledge" expresses or embodies.

The word and its meaning

Philosophers who proceed in this way draw a distinction between a word and its meaning, and a meaning is generally considered to be the concept which that particular word has or expresses. It is usually further assumed that though concepts are not identical with words, that is, with linguistic expressions, language is the medium in which the meaning of such concepts is displayed or expressed.

The investigation into the nature of knowledge often begins in a similar fashion with the study of the use of the word "knowledge" and of certain cognate expressions and phrases found in everyday language. A survey of such locutions reveals important differences in their uses: one finds such expressions as "know him," "know that," "know how," "know where," "know why," or "know whether." These differences have been explored in detail, especially in the 20th century. The expression "know *x,*" where "*x*" can be replaced by a proper name, as in "I know Jones" or "He knows Rome," has been taken by some philosophers, notably Bertrand Russell (1872–1970), to be a case of knowledge by acquaintance. Russell thought its characteristic use was to express the kind of knowledge one has when one has first-hand familiarity with a certain object, person, or place. Thus, one could not properly say in the 20th century, "I know Julius Caesar," since this would imply that one had met or was directly acquainted with a person who had died some 2,000 years ago. This sense or use of "know" becomes important in the theory of perception and in sense-data theory, since some philosophers, such as Russell and G.E. Moore (1873–1958), have held that one's awareness of a sense-datum (a notion to be discussed later) is a case of direct acquaintance, whereas one's acquaintance with a physical object, such as a human hand, is not.

The phrases "know that" and "know how" have also played fundamental roles in the theory of knowledge. The British philosopher Gilbert Ryle (1900–76), for instance, argued that "know how" is normally used to refer to a kind of skill that a person has, such as knowing how to swim. One could have such knowledge without being able to explain to another what it is that one knows in such a case, that is, without being able to convey to another the knowledge required for that person to develop the same skill. "Know that," in contrast, does not seem to denote the possession of a skill or aptitude but rather the possession of specific pieces of information, and the person who has knowledge of this sort can generally convey it to others. To know that the Concordat of Worms was signed in the year 1122 would be an example of this sort of knowledge. Ryle has argued that, given these differences, some cases of knowing how cannot be reduced to cases of knowing that and, accordingly, that the kinds of knowledge expressed by these phrases are independent of one another.

In general, the philosophical tradition from the Greeks to the present has focused on the kind of knowledge expressed when it is said that someone knows that such and such is the case, *e.g.,* that A knows that snow is white. This sort of knowledge, called propositional knowledge, raises the classical epistemological questions about the truth or falsity of the asserted claim, the evidence for it, and a host of other problems. Among them is the much debated issue of what kind of thing is known when one knows that *p, i.e.,* what counts as a substitution instance of *p.* The list of such candidates includes beliefs, propositions, statements, sentences, and utterances of sentences. Each has or has had its proponents, and the arguments pro and con are too subtle to be explored here. Two things should, however, be noted in this connection: first, that the issue is closely related to the problem of universals (*i.e.,* whether what is known to be true is an abstract entity, such as a proposition, or whether it is a linguistic expression, such as a sentence or a sentence-token) and, second, that it is agreed by all sides that one cannot have knowledge, in this sense of "knowledge," of that which is not true. One of the necessary conditions for saying that A knows that *p* is that *p* must be true, and this condition can therefore be regarded as one of the main elements in any accurate characterization of knowledge.

Propositional knowledge

SIX DISTINCTIONS OF KNOWLEDGE

Mental versus nonmental conceptions of knowledge. Philosophers have asked whether knowledge is a state of mind, *i.e.,* a special kind of awareness of things. That it is has been argued by philosophers since at least the 5th century BC. In *The Republic* Plato provided the first extensive account of such a view. He regarded knowing as a mental faculty, akin to but different from believing or opining. Contemporary versions of this sort of theory regard knowing as one member of a sequence of mental states that involve increasing certitude. This spectrum would begin with guessing or conjecturing at the lowest end of certitude, would include thinking, believing, and feeling sure as expressing stronger attitudes of conviction, and would end with knowledge as the highest of all these states of mind. Knowledge, in all views of this type, is a form of consciousness, the strongest degree of awareness humans possess, and accordingly it is common for proponents of such views to hold that, if A knows that *p,* A must be conscious of what he knows. This view is normally expressed by saying that, if A knows that *p,* A knows that he knows that *p.*

Knowledge as a form of consciousness

Many 20th-century philosophers have rejected the notion that knowledge is a mental state. In *On Certainty* (1969) Ludwig Wittgenstein says: " 'Knowledge' and certainty belong to different *categories.* They are not two mental states like, say surmising and being sure." But, if knowing is not a mental state, then what is it? These philosophers have accepted the challenge of trying to give a different characterization of what it means to say that a person knows something. They typically begin by pointing out that a person can know that *p* without knowing that he knows it (a good example is in fact to be found in Plato's *Meno,* where Socrates gradually elicits from a slave boy geometrical knowledge that the boy was not aware he had). They then proceed to argue that it is a mistake to assimilate cases of knowing to cases of doubting, feeling a pain, or having a certain opinion about something. All of these latter are mental states, and they are such that a person who has such a state is aware that he does.

These philosophers, moreover, typically deny that knowing can be described as being a single thing, such as a state of consciousness. Instead, they claim that one can ascribe knowledge to someone, or to oneself, when certain complex conditions are satisfied, among them certain behavioral conditions. For example, if a person can always give the right answers to questions under test conditions, one would be entitled to say that the person has knowledge of the issues under consideration. Knowing on this account seems tied to the capacity to perform in certain ways under certain standard conditions. Accordingly, though such performances may involve the exercise of intelligence or other mental factors, the attribution of knowledge to someone is not merely the attribution of a certain mental state or state of awareness to that person (as seen in the case of the slave boy in the *Meno*).

A well-known variant of such a view was advanced by J.L. Austin in his 1946 paper "Other Minds." Austin claimed that, when one says "I know," one is not describing anything, let alone one's psychology or a mental state. Instead, one is engaging in a social act, *i.e.,* one is indicating that one is in the position (has the credentials and the reasons) to assert *p* in circumstances where it is necessary to resolve a doubt. When these conditions are satisfied, one can correctly be said to know.

Occurrent versus dispositional conceptions of knowledge. A distinction closely related to the previous one is that between occurrent and dispositional conceptions of knowledge. The difference between occurrences and dispositions can be illustrated with respect to sugar. A sugar cube will dissolve if put into water. One can thus say that, even if the cube is not now dissolving as it sits on the table, it will do so under certain conditions. This propensity to dissolve is what is meant by a disposition, and it is a feature sugar has at all times and in all conditions. It can be contrasted with sugar's actually dissolving when immersed in liquid, which is an occurrence, that is, an event happening at a specific place and at a specific time.

These terms also apply to mental events. One can say of Smith, who is working on a problem, that he has just seen the solution. According to this way of speaking, there is a certain answer that Smith is presently aware of and to which he is attending. In such a case Smith's knowledge is occurrent. But one can also ascribe a different sort of knowledge to Smith. Though Smith is perhaps not now thinking of his home address, he certainly knows it in the sense that, if he were asked, he could produce the correct answer. One can thus have knowledge that one is not aware of at a given moment. One can thus say, as with sugar, that knowledge may be either occurrent or dispositional in character, *i.e.,* that one may or may not be in an immediate state of self-awareness with respect to *p,* but that in either case it can be said that the person knows that *p.*

It should be noted that the distinction between dispositional and occurrent knowledge thus applies to cases of "knowing that" as well as to cases of "knowing how" and thus is a powerful conceptual tool for analyzing different sorts of epistemic notions. The concept of a disposition has itself been further analyzed, for example by Roderick M. Chisholm (b. 1916), in counterfactual terms, and it has been proposed by many philosophers that the knowledge expressed by causal laws (laws of nature) is counterfactual and thus dispositional in character.

Counterfactual knowledge

A priori versus a posteriori knowledge. A sharp distinction has been drawn since at least the 17th century between two types of knowledge: a priori knowledge and a posteriori knowledge. The distinction plays an especially important role in the philosophies of David Hume (1711–76) and Immanuel Kant (1724–1804). It is also found in many contemporary, empirically oriented theories of knowledge, which typically hold that all knowledge about matters of fact derives from experience and is therefore a posteriori and that in consequence such knowledge is never certain but at most only probable.

The difference between these types of knowledge is easy to illustrate by means of examples. Consider the sentences "All husbands are married" and "All Model-T Fords are black" and assume that both statements are true. But how does one come to know that they are true? In the case of the first, the answer is that, if one thinks about the meaning of the various words in the sentence, one can see that the sentence is true. One can see that this is so because what is meant by "husband" is the same as what is meant by "married male." Thus, by definition, every husband is a married male, and, accordingly, every husband is married. In calling such knowledge a priori, philosophers are pointing out that one does not have to engage in a factual or empirical inquiry in order to determine whether the sentence is true or not. One can know this merely on the basis of reflection and thus prior to or before any investigation of the facts.

In contrast, the second statement can be determined to be true only after such an investigation. One may well know that the Model-T Ford was an automobile built prior to World War II and accordingly would understand what all the words in the sentence mean. Nonetheless, understanding alone would not be sufficient to allow one to determine whether the sentence is true or not. Instead, some kind of empirical investigation is required in order to arrive at such a judgment; the knowledge thus acquired is a posteriori, or knowledge after the fact.

There are sets of distinctions related to the one just developed and in terms of which the two propositions can also be differentiated. They are necessary versus contingent, analytic versus synthetic, tautological versus significant, and logical versus factual.

Necessary versus contingent propositions. A proposition is said to be necessary if it holds (is true) under all possible circumstances or conditions. "All husbands are married" is such a proposition. There are no possible or conceivable conditions under which this statement would not be true (on the assumption, of course, that the words "husband" and "married" are taken to mean what they ordinarily mean). In contrast, "All Model-T Fords are black" holds in some circumstances (those actually obtaining, and that is why the proposition is true), but it is easy to imagine circumstances in which it would not be true—for instance,

if somebody painted one of those cars a different colour. To say, therefore, that a proposition is contingent is to say that it holds in some but not in all possible circumstances. Some necessary propositions, such as "All husbands are married" are a priori (though not all are) and most contingent propositions are a posteriori.

Analytic versus synthetic propositions. A proposition is often said to be analytic if the meaning of the predicate term is contained in the meaning of the subject term. Thus, "All husbands are married" is analytic because the term "husband" includes as part of its meaning "being married." A term is said to be synthetic if this is not so. Therefore, "All Model-T Fords are black" is synthetic since the term "black" is not included in the meaning of "Model-T Ford." Some analytic propositions are a priori, and most synthetic propositions are a posteriori. These distinctions were used by Kant to ask one of the most important questions in the history of epistemology, namely, whether a priori synthetic judgments are possible (see below for a discussion of this question).

Tautological versus significant propositions. A proposition is said to be tautological if its constituent terms repeat themselves or if they can be reduced to terms that do, so that the proposition is of the form "a = a." In such a case the proposition is said to be trivial and empty of cognitive import. A proposition is said to be significant if its constituent terms are such that the proposition does provide new information about the world. It is generally agreed that no significant propositions can be derived from tautologies. One of the objections to the ontological argument is that no existential (significant) proposition can be derived from the tautological definition of "God" with which the argument begins. Tautologies are generally known to be true a priori, are necessary, and are analytic; and significant statements are generally a posteriori, contingent, and synthetic.

In the ontological argument, for example, God is defined (roughly speaking) as the only perfect being. It is then argued that no being can be perfect unless it exists; therefore, God exists. But, as Hume and Kant pointed out, it is fallacious to derive a factual statement about the existence of God from the definition of God as a perfect being (see the discussion of St. Anselm below).

Logical versus factual propositions. The term "logical" in this connection is used in a wide sense to include a proposition such as "All husbands are married." By analyzing the meaning of its constituent terms one can reduce the proposition to a logical truth, *e.g.*, to "A and B implies A." In contrast, factual propositions, such as "All Model-T Fords are black," have syntactical and semantic structures that differentiate them from any propositions belonging to logic, even in the broad sense mentioned above. The theorems of logic are often a priori (though not always), are always necessary, and are typically analytic. Factual propositions are generally a posteriori, contingent, and synthetic.

These various distinctions are widely appealed to in present-day philosophy. For instance, Saul Kripke (b. 1941) in "Naming and Necessity" (1972) has used these notions in an effort to solve a long-standing problem, namely, how true identity statements can be nontrivial. The problem, first articulated by Gottlob Frege in "On Sense and Reference" (1893) and later independently addressed by Russell, begins with the assumption that the sentences "Scott is Scott" and "Scott is the author of *Waverley*" are both identity sentences and are true and that the former is trivial while the latter is not. The puzzle arises from the further assumption that any true identity sentence simply says of some object that it is identical with itself. Hence, all such sentences should be trivial. Clearly, however, "Scott is the author of *Waverley*" is not trivial. But, if it is not, how is this possible?

Trivial and nontrivial identity sentences

Kripke argues that all true identity sentences are necessary (*i.e.*, that they hold in all possible worlds) and that some of these, such as "Scott is Scott," are known a priori and accordingly are trivial; but, he argues, some true identity sentences are not known a priori but only a posteriori and are not trivial. In cases of the latter sort, their nontriviality is a function of their being known to be true only after some sort of inquiry or investigation. It is the investigation that provides new information.

A good example would be the following. At one time in human history, ancient peoples did not know that what they called "the evening star" was the same planet called "the morning star." But eventually the Babylonians discovered through astronomical observation that the morning star is the planet Venus as it appears in the morning sky and that the evening star is the planet Venus as it appears in the evening sky. The discovery that these two appearances are appearances of the same object amounted to discovering more than that Venus is Venus. It provided new information, and that is why "the morning star is identical with the evening star" is significant in a way in which "Venus is Venus" is not, even though all of the descriptive terms in both sentences refer to exactly the same object. In similar fashion, the a posteriori finding that it was Scott who wrote *Waverley* explains the nontriviality of "Scott is the author of *Waverley*." But no such investigation was needed to determine that Scott is Scott.

Knowledge by acquaintance and knowledge by description. The distinction between knowledge by acquaintance and knowledge by description was introduced by Bertrand Russell in connection with his celebrated theory of descriptions. Here only the epistemological (as distinct from the logical) version of his theory will be considered. It was invented by Russell to lend support to the basic thesis of empiricism that all knowledge of matters of fact (*i.e.*, all a posteriori knowledge) derives from experience. Russell's program is both reductive and foundationalist. It tries to show that man's system of knowledge is stratified: that some types of knowledge depend on others but that some do not and that the latter form the foundational units which give support to the whole epistemic system. He argued that, because these basic units rest upon direct experience, ultimately all factual knowledge is derivable from experience.

Russell's argument begins with a distinction between two different types of knowledge, that which is and that which is not based on direct experience. Nearly all of man's knowledge is of the latter type. For example, it is known that some 2,000 years ago there lived a Roman statesman named Augustus, that he was the successor to Julius Caesar, who had been assassinated, and that he was a friend of the historian Livy. But, since none of these pieces of information is presently known on the basis of personal experience, what justification is there for calling them instances of knowledge?

Russell argued that information based on direct experience is basic and needs no justification; he called it "knowledge by acquaintance." Information not based on direct experience he called "knowledge by description." One is justified in calling such information knowledge, if one can show that it can be traced back to and thus ultimately rests upon knowledge by acquaintance. To show how this is so in a particular case is to legitimate that particular piece of information as a specimen of knowing. Here is how this reductive process would work in the case of what is known about Augustus.

Whatever information people in the 20th century have about Augustus probably comes to them from literary works, such as Livy's history of Rome. Such information thus comes secondhand, via descriptions in books about the life and activities of Augustus. But why call such descriptions knowledge? The answer is that through a historical process one can trace such information back to an original source like Livy, who was a contemporary of Augustus. One learns, via this process, that Livy in his history of Rome is reporting events that he had witnessed himself or that he had learned from other eyewitnesses. One can call what he tells about Augustus knowledge, because it is testimony that is based upon his or someone else's direct experience. Thus, knowledge by description is a legitimate form of knowledge, even though it is ultimately dependent upon knowledge by acquaintance.

Russell's reductive thesis then was that all legitimate specimens of knowledge are either based upon direct experience or can be shown to be dependent upon such direct experience via a chain of tight historical or causal links.

His theory was therefore a form of empiricism, because it tried to show how all knowledge of matters of fact could be derived from experience.

But there is a further feature of the theory, stemming from the empirical tradition of John Locke (1632–1704) and David Hume, that gives a special twist to the notion of "knowledge by acquaintance." According to this tradition, knowledge by acquaintance is always knowledge based upon what Hume called "impressions," or upon what Russell called "sense-data." These for Russell were mental entities that generally, but not always, reflected the characteristics actually possessed by physical objects. But, unlike physical objects, sense-data were the objects directly apprehended in an act of perception. What Russell meant by "direct apprehension" or "direct perception" was itself explicated in terms of the concepts of inference and non-inference. He held that direct perception, *i.e.,* the perceptual awareness of a sense-datum, involves no inference and, accordingly, that knowledge by acquaintance is identical with the perception of sense-data.

Sense-data

The difference between inferential and noninferential perception can be illustrated by an example. Suppose one is working in a room and hears a sound that emanates from an outside source. (Russell considered hearing to be a form of perception.) In such a case the sound is a sense-datum. One need not infer that one is hearing a sound; there is a direct awareness of it. This would be a case of knowledge by acquaintance. On the basis of what one hears in this direct fashion, one might then infer (guess, conjecture, hypothesize) that what is causing the sound is a motorcycle located outside of the room, something that one who is in the room cannot see directly. If one is correct in this supposition, the information obtained in this way would be a case of indirect knowledge. In such a case, one's knowledge that there is a motorcycle in the street is dependent on (and in Russell's sense, reducible to) one's direct awareness of a sound. The example illustrates how indirect knowledge, such as knowledge by description, is derived from direct knowledge, such as knowledge by acquaintance, and in turn how this latter depends upon the direct awareness of sense-data.

It should be mentioned that the distinction between knowledge by acquaintance and knowledge by description can be defended as legitimate and useful independently of a commitment to sense-data theory. In Russell's work the objects of direct awareness are sense-data, but sense-data theory today has few proponents. A philosopher thus might hold that one at least sometimes directly perceives physical objects (which are not sense-data) while accepting that one's knowledge of past events and persons is indirect and is thus knowledge by description.

Description versus justification. Epistemology during its long history has engaged in two different sorts of tasks. One of these is descriptive in character. It aims to depict accurately certain features of the world, including the contents of the human mind, and to determine whether these should count as specimens of knowledge. A philosophical system with this orientation is, for example, the phenomenology of Edmund Husserl (1859–1938). Husserl's aim was to give an exact description of the notion of intentionality, which he characterized as consisting of a certain kind of "directedness" toward an object. Suppose the object is an ambiguous drawing, such as the duck/rabbit sketch found in the *Philosophical Investigations* of Ludwig Wittgenstein. A person looking at the sketch is not sure whether it is a drawing of a duck or of a rabbit. Husserl claimed that the light rays reaching the eye from such an ambiguous drawing are identical whether one sees the image of a duck or of a rabbit and that the difference in perception is due to the viewer's structuring of what he sees in the two cases. The theory tries to describe how such structuring takes place, and it ultimately becomes very complex in the account it gives.

In a famous passage in *Philosophical Investigations* (1953), Wittgenstein states that "explanation must be replaced by description," and much of his work was devoted to carrying out that task, as, for example, in his account of what it is to follow a rule. Another example of descriptive epistemology is found in the writings of such sense-data theorists as Moore, Price, and Russell. They begin with the question of whether there are basic apprehensions of the world, free from any form of inference, and in those cases where they have argued that the answer is yes, they have tried to describe what these are and why they should count as instances of knowledge. Russell's thesis that the whole edifice of knowledge is built up from a foundation composed of ingredients with which human beings are directly acquainted illustrates the close connection between the attempt to characterize various types of knowledge and this descriptive endeavour. The search by some logical positivists, such as Moritz Schlick (1882–1936), Otto Neurath (1882–1945), and A.J. Ayer (b. 1910) for protocol sentences, sentences that describe what is given in experience without inference, is a closely related example of this kind of descriptive practice.

Epistemology has a second function, which, in contrast to the descriptive one, is justificatory or normative. Philosophers concerned with this function start from the fact that all human beings have beliefs about the world, some of which are erroneous and some of which are not. The question to them is how one can justify (defend, support, or provide evidence for) certain sets of beliefs. The question has a normative import since it asks, in effect, what one ideally ought to believe. (In this respect epistemology has close parallels to ethics, where normative questions about how one ought ideally to act are asked.) This approach quickly takes one into the central domains of epistemology. It raises such questions as: Is knowledge identical with justified true belief? Is the relationship between evidence for a belief and the belief itself a probability function? If not, what is it? What indeed is meant by "justification" and what sorts of conditions have to be satisfied before one is entitled to say that a belief or set of beliefs is justified? These two differing aspects of epistemology are not inconsistent and indeed are often found intertwined in the writings of contemporary philosophers.

Normative function of epistemology

Knowledge and certainty. The relationship between knowledge and certainty is complex, and there is considerable disagreement about the matter. Are these concepts the same? If not, how do they differ? Is it possible for someone to know that p without being certain that p? Is it possible for someone to be certain that p without knowing that p? These are the central issues around which the debate revolves. The various answers that have been proffered depend on how the concepts of knowledge and certainty are analyzed. If one holds, for instance, that knowing is not a psychological state but that certainty is, then one would deny that the concepts are identical. But if one holds that knowing represents the highest degree of assurance which humans can obtain with respect to the truth of p, and that such a maximal degree of assurance is a psychological state, one will interpret the concepts to be equivalent. There have been proponents on both sides of this issue.

Further complicating the discussion are subtle distinctions drawn by 20th-century philosophers. For instance, in "Certainty" (1941) G.E. Moore claimed that there are four main types of idioms in which the word "certain" is commonly used: "I feel certain that," "I am certain that," "I know for certain that," and finally "It *is* certain that." He points out that "I feel certain that p" may be true when p is not true but that there is at least one use of "I know for certain that p" and "It is certain that p" which is such that neither of these sentences can be true unless p is true. Moore argues that it would be self-contradictory to say "I knew for certain that he would come but he didn't," whereas it would not be self-contradictory to say "I felt certain he would come but he didn't." In the former case, the fact that he did not come proves that one did not know that he would come, but, in the latter, the fact that he did not come does not prove that one did not feel certain he would. "I am certain that" differs from "I know for certain that" in allowing the substitution of the word "sure" for the word "certain." One can say "I feel sure (rather than certain)" without a change of meaning, whereas in "I know for certain" or "It is certain that" this substitution is not possible. On the basis of these sorts of considerations Moore contends that "a thing can't be certain unless it

is *known.*" He states that this is what distinguishes the word "certain" from the word "true." A thing that nobody knows may well be true, but it cannot possibly be certain. He thus infers that a necessary condition for the truth of "It is certain that *p*" is that somebody should know that *p* is true. Moore is therefore one of the philosophers who answers in the negative the question of whether it is possible for *p* to be certain without being known.

Moore also argues that to say "Someone knows that *p* is true" cannot be a sufficient condition for "It is certain that *p.*" If it were, it would follow that, in any case in which at least someone did know that *p* was true, it would always be false for anyone to say "It is not certain that *p*"; but clearly this is not so. If one person says that it is not certain that Smith is still alive, he is not thereby committing himself to the statement that nobody knows that Smith is still alive: the speaker's statement is consistent with Smith's still being alive, and both he himself and other persons know this. Moore is thus among those philosophers who would answer in the negative the question of whether the concepts of knowledge and certainty are the same. Though it is widely accepted that to affirm that somebody knows that *p* implies that somebody is certain that *p,* the case of the slave boy in Plato's *Meno* seems, at least at first glance, to be a counterinstance. Meno may know, in a dispositional sense, certain theorems of geometry without knowing that he knows, and, if he does not know that he knows, then it would seem that he cannot be certain that he does know. But it has also been argued that, once his disposition to know has been actualized and his knowledge has become occurrent, then, insofar as he does know in this occurrent sense, he is certain of what he knows.

The most radical position on these matters is to be found in Wittgenstein's *On Certainty,* published posthumously in 1969. Wittgenstein holds that knowledge is radically different from certitude and that neither concept entails the other. It is thus possible to be in a state of knowledge without being certain and to be certain without having knowledge. As he writes: "Instead of 'I know' . . . couldn't Moore have said: 'It stands fast for me that . . .' ? and further: 'It stands fast for me and many others. . . .' "

Standing fast

"Standing fast" is one of the terms Wittgenstein uses for certitude and is to be distinguished from knowing. For him certainty is to be identified with acting, not with seeing propositions to be true, the kind of seeing that issues in knowledge. As he says: "Giving grounds, justifying the evidence comes to an end—but the end is not certain propositions striking us immediately as true—*i.e.,* it is not a kind of *seeing* on our part; it is our *acting* which lies at the bottom of the language game."

ORIGINS OF KNOWLEDGE

Philosophers not only wish to know what knowledge is but also how it originates. This motivation is based, at least in part, on the supposition that an investigation into the provenance of knowledge can help cast light on its nature. From the time of the Greeks to the present, therefore, one of the major themes of epistemology has been a quest into the sources of knowledge.

Plato's *The Republic* contains one of the earliest systematic arguments to the effect that sense experience cannot be a source of knowledge. The argument begins with the assertion that ordinary persons have a clear grasp of certain concepts, that of equality, for instance. In other words, people know what it means to say that A and B are equal, no matter what A and B are. But where does such knowledge come from? One may wonder, for instance, whether it is provided by vision and consider the claim that two pieces of wood are of equal length. A close inspection of these pieces of wood, however, shows them to differ slightly, and the more detailed the inspection, via various degrees of magnification, the more disparity one notices. It follows that visual experience cannot be the fount of the concept of equality. Plato applies this result to the operations of all the five senses and concludes that sense experience in general cannot be the origin of such knowledge. It must therefore have another source, which he regards as prenatal (one such account is found in the myth of Er in Book X).

The mathematical example Plato selects to illustrate that the origin of knowledge is not in sense experience is highly significant; indeed it is one of the signs of his perspicacity that he should pick such an example. For, as the subsequent history of philosophy reveals, the strongest case for the notion that at least some knowledge does not derive from sense experience lies in mathematics. Mathematical entities are abstractions—perfect triangles, disembodied surfaces and edges, lines without thickness, and extensionless points—and none of these exists in the physical world, *i.e.,* the world apprehended by the senses. It might be thought that, had Plato selected a different example, say, the colour red, his argument would have been less convincing. But it is a further sign of his genius that he discusses colours as well as mathematical notions and provides good reasons for holding that seeing examples or specimens of red (or any other colour) is not equivalent to knowing what that colour is. Such knowledge must therefore have a different genesis than sense experience.

Innate versus learned. The puzzle about origins of knowledge has led historically to two different kinds of issues. One of these is the question of whether knowledge (or at least certain kinds of knowledge) is innate, meaning that it is not acquired or learned through experience but in some important sense is present in the human psyche at birth. The matter is still a live issue today, not only in philosophy but also in linguistics and psychology. The linguist Noam Chomsky (b. 1928), for example, has asserted that the "projection phenomenon"—the ability of children to construct sentences that they have never heard before and that are grammatical—is proof of inherent conceptual structures, whereas the experimental psychologist B.F. Skinner (b. 1904) has tried to show that all knowledge is the product of learning through environmental conditioning by means of the processes of reinforcement and reward.

In the extensive historical literature on this topic both the notion of "innateness" and that of "learning" have been given various interpretations. Sometimes, for instance, innateness carries only the sense of a disposition or propensity, but in stronger versions of the thesis, such as Plato's, it is affirmed that humans possess actual pieces of prenatal knowledge. "Learning" also is given a variety of meanings, ranging from trial-and-error methods to inexplicit types of "absorption" of information. There are also a range of "compromise" theories. These typically claim that humans have some knowledge that is innate—the awareness of God, the principles of moral rightness and wrongness, and certain mathematical theorems being favoured examples—whereas other kinds of knowledge—such as knowledge by acquaintance—are gained through experience.

Interpretations of innateness and learning

Rationalism versus empiricism. The second issue that emerges from considerations of the origins of knowledge focuses on the distinction between rationalism and empiricism. Though closely related to the issue of innateness versus learning, the question in this case concerns the nature of the source from which knowledge arises. The history of discussion of the issue indicates that two main sources have been identified and argued for: reason and experience.

Rationalism is the thesis that the ultimate source of knowledge is to be found in human reason. What reason is, in turn, is a difficult question. But, generally speaking, it is assumed that reason is a feature of the human mind that differs not just in degree but in kind from bodily sensations, feelings, and certain psychological attitudes, such as disgust or enthusiasm. For some writers, such as Plato, reason is a faculty, a special facility or structure of the mind. Many later philosophers reject any sort of faculty psychology, and some of them tend to interpret reason in dispositional or behavioral ways. But, whatever the interpretation, a rationalist must hold that reason has a special power for grasping reality. It is the exercise of reason that allows human beings to understand the world they live in. Such a thesis is double-sided: it holds, on the one hand, that reality is in principle knowable and, on the other hand, that there are human, distinctively mental, powers capable of apprehending it. One thus might define

rationalism as the theory that there is an isomorphism (a mirroring relationship) between reason and reality which makes it possible for the former to apprehend the latter just as it is. Rationalists affirm that, if such a correspondence were lacking, the effort of human intelligence to understand the world would be impossible.

Empiricism is often defined as the doctrine that all knowledge comes from experience. Almost no philosopher, however, has ever literally held that all knowledge comes from experience. Locke, who is the empiricist par excellence, thought there is some knowledge human beings have—which he calls "trifling ideas" (or trivialities), such as $a = a$—that does not derive from experience; but he regarded such knowledge as empty of content. Hume held similar views.

Empiricism thus generally allows for a priori knowledge while denigrating its significance, and accordingly it is more accurate to define it as the theory that all knowledge about matters of fact derives from experience. When defined in this way empiricism does represent a significant contrast to rationalism. Rationalists hold that human beings have knowledge about matters of fact which is anterior to experience and yet which does tell them something significant about the world and its various features. Empiricists would deny that this is possible.

The meaning of the term experience is generally limited to the impressions and sensations received by the senses. Thus, knowledge is the information apprehended by the five sense modalities—hearing, seeing, touching, tasting, and smelling. Such knowledge is always about matters of fact, about what one can see, touch, hear, taste, or smell. For strict empiricists this definition has the implication that the human mind is passive—a tabula rasa, in Locke's idiom; it is an organ that receives impressions and more or less records them as they are. This conception of the mind has seemed counterintuitive to many philosophers, especially those in the Kantian tradition. But it also poses serious challenges for empiricists. For example, it raises the question of how one can have knowledge of items, such as a dragon, that cannot be found in experience.

In response, the classical empiricists such as Locke and Hume have tried to show how the complex concept of a dragon can be reduced to simple concepts (such as wings, the body of a snake, the head of a horse), all of which derive from direct impressions of such items. On such a view the mind is still considered to be primarily passive, but it is conceded that it has some active functions, such as being able to combine simple impressions and ideas into complex ideas.

Challenges to empiricism

There are further difficulties: the empiricist must explain how abstract ideas, such as the concept of a perfect triangle, can be reduced to elements apprehended by the senses when no perfect triangles are actually found in nature, and he must also give an account of how general notions are possible. It is obvious that one does not experience "mankind," but only particular individuals, through the senses; yet such general notions are meaningful, and propositions containing these concepts are known to be true. The same difficulty applies to colour concepts. Some empiricists have argued that one arrives at the concept of red, for example, by abstracting from individual items that are red. But the difficulty with this suggestion is that one would not know what to count as an instance of red unless one already had such a concept in mind; and, if that is so, it would seem that experience cannot be the source of the concept. It is generally felt that, despite ingenious attempts by empiricists to deal with such issues, their solutions have not been wholly successful. Indeed, the history of epistemology has to a large extent been a dialectic between rationalism and empiricism in an effort to meet skeptical challenges that are designed to undermine both positions.

SKEPTICISM

Many philosophers past and present and many nonphilosophers who are studying philosophy for the first time have been struck by the seemingly indecisive nature of philosophical argumentation. For every argument, there seems to be a counterargument; and for every position, a counterposition. To a considerable extent skepticism is born of such reflection. Some of the ancient skeptics contended, for example, that all arguments are equally bad and, accordingly, that nothing can be proved. The American philosopher Benson Mates claims to be a modern representative of this tradition, except that he believes all philosophical arguments to be equally good. But he insists that, because they are, they invariably issue in conceptual deadlocks and resolve nothing.

Ironically, skepticism is itself a type of philosophy, and the question has been raised whether it manages to escape its own demurrers. Does it offer arguments, and, if so, are they decisive? The answers to these questions depend on what is meant by skepticism. Historically, the term refers to a complex set of practices taking many different forms—from stating explicit theories to assuming negative attitudes without much propositional content. Thus, it is difficult to define. But, however it is understood, skepticism represents a set of challenges to the claim that human beings do possess or can acquire knowledge.

In giving even this minimal characterization, it is important to emphasize that both dogmatists and skeptics accept a definition of knowledge that implies two things: that, if a person, A, knows that *p*, then *p* is true and that, if a person, A, knows that *p*, then A cannot be mistaken, meaning that it is logically impossible that A could be wrong. If a person says that he knows Smith will arrive at 9:00 AM, and Smith is not there at 9:00 AM, then that person would have to withdraw the claim to know. He might say instead that he thought he knew or that he felt sure. But he could not continue rationally to insist that he knew if what he claimed to know turned out to be false.

It should also be stressed that, given this definition of knowledge, the skeptic does not have to show that A is actually mistaken in claiming to know that *p*. All he has to show is that it is possible that A might be mistaken. Hence arises the skeptic's practice of searching for a possible counterexample to a claim. If A states that he has had a certain experience, for instance, that of having personally spoken with Smith, who assured him he would keep his appointment at 9:00 AM, then the skeptic can point out that, although one could have such an experience, it is still possible that Smith might not show up; and, if so, A's claim to know is untenable. In effect, by emphasizing the notion of possibility, the skeptic is pointing out that there is a logical gap between the criteria that support the claim and the claim itself. The criteria might be satisfied, and yet the claim might be false; but, if such a possibility exists, the original assertion cannot be a specimen of knowing.

More generally, radical skepticism has tried to show that one might (*i.e.*, could possibly) have all the experiences associated with normal perception or behaviour and yet be wholly mistaken in thinking that these experiences correlate with anything in the external world. For example, a brain in a vat might be programmed by scientists to have the sensation of seeing a tree, even though it is not in fact seeing a tree. Thus, there is a gap between the experience the brain is having and external reality; accordingly, its claim to know on the basis of such a visual experience is mistaken. The skeptic's point is that the disparity between external reality and felt experience is always possible and, accordingly, that knowledge claims based upon such experience cannot be defended.

The gap between external reality and felt experience

The ability to find counterexamples explains why skeptics do not challenge but indeed accept the dogmatist's definition of knowledge. That they do so is important because it means that they are not arguing at cross-purposes with their opponents. What they challenge is not the meaning of knowledge but the contention that anybody actually has knowledge in that sense.

Nearly all of the major epistemological theories of philosophy have given rise to skeptical reactions. Many of the greatest thinkers in the Western tradition have assumed that by means of reason or sense experience one can come to have knowledge of reality. But skepticism has challenged the validity of both of these appeals. Skeptics have developed wholesale arguments to undermine the efforts to show that reason and sense experience, which seem to be the only possible candidates, are reliable sources

of knowledge. Descartes, for example, considered the hypothesis that an evil genius may delude people into thinking that they are experiencing the real world when they are not. With regard to major epistemological problems, such as the "other-minds problem," the problem of memory, the problem of induction, and the problem of self-knowledge, skeptical doubts have challenged the validity of reason and of sense experience and thus of claims to have knowledge of various aspects of reality. How some of these moves and countermoves actually take place are addressed below. (AV.S.)

The history of epistemology

ANCIENT PHILOSOPHY

Pre-Socratics. The central focus of ancient Greek philosophy was its attempt to solve the problem of motion. Many pre-Socratic philosophers thought that no logically coherent account of motion and change could be given. This problem was a concern of metaphysics, not epistemology, however, and in the present context it suffices merely to allude to the arguments of Parmenides and Zeno of Elea against the possibility that anything moves or changes. The consequence of this position for epistemology was that all major Greek philosophers held that knowledge must not itself change or be changeable in any respect. This requirement motivated Parmenides, for example, to hold that thinking is identical with being (what exists or is unchanging) and that it is impossible to think of "nonbeing" or "becoming" (what changes) in any way.

Plato. Plato (*c.* 427–347 BC) accepted the Parmenidean constraint on any theory of knowledge that both knowledge and its objects must be unchanging. One consequence of this, as Plato pointed out in *Theaetetus,* is that knowledge cannot have physical reality as its object. In particular, since sensation and perception have various kinds of motions as their objects, knowledge cannot be the same as sensation or perception. The negative thesis of Plato's epistemology consists, then, in the denial that sense experience can be a source of knowledge on the ground that the objects apprehended through the senses are subject to change. To the extent that humans have knowledge, they attain it by transcending the information provided by the senses in order to discover unchanging objects. But this can be done only by the exercise of reason, and in particular by the application of the dialectical method of inquiry inherited from Socrates.

The Platonic theory of knowledge is thus divided into two parts: a quest first to discover whether there are any unchanging objects and to identify and describe them and second to illustrate how they could be known by the use of reason, that is, via the dialectical method. Plato used various literary devices for illustrating his theory; the most famous of these is the allegory of the cave in Book VII of *The Republic.* The allegory depicts ordinary people as living locked in a cave, which represents the world of sense-experience; in the cave people see only unreal objects, shadows, or images. But through a painful process, which involves the rejection and overcoming of the familiar sensible world, they begin an ascent out of the cave into reality; this process is the analogue of the application of the dialectical method, which allows one to apprehend unchanging objects and thus acquire knowledge. In the allegory, this upward process, which not everyone is competent to engage in, culminates in the direct vision of the sun, which represents the source of knowledge.

In searching for unchanging objects, Plato begins his quest by pointing out that every faculty in the human mind apprehends a set of unique objects: hearing apprehends sounds but not odours; the sense of smell apprehends odours but not visual images; and so forth. Knowing is also a mental faculty, and therefore there must be objects that it apprehends. These have to be unchanging, whatever they are. Plato's discovery is that there are such entities. Roughly, they are the items denoted by predicate terms in language: such words as "good," "white," or "triangle." To say "This is a triangle" is to attribute a certain property, that of being a triangle, to a certain spatiotemporal object, such as a particular figure drawn on a blackboard. Plato is here distinguishing between specific triangles that can be drawn, sketched, or painted and the common property they share, that of being triangular. Objects of the former kind he calls particulars. They are always located somewhere in the space-time order, that is, in the world of appearance. But such particular things are different from the common property they share. That is, if x is a triangle, and y is a triangle, and z is a triangle, x, y, and z are particulars that share a common property, triangularity. That common property is what Plato calls a "form" or "idea" (not using this latter term in any psychological sense). Unlike particulars, forms do not exist in the space-time order. Moreover, they do not change. They are thus the objects that one must apprehend in order to acquire knowledge.

Similar remarks apply, for example, to goodness, whiteness, or being to the right of. Particular things change; they come into and go out of existence. But whiteness never changes, and neither does triangularity; and, if they do not change, they are not subject to the ravages of time. In that sense, they are eternal.

Dialectical method

The use of reason for discovering unchanging forms is exercised in the dialectical method. The method is one of question and answer, designed to elicit a real definition. By a "real definition" is meant a set of necessary and sufficient conditions that exactly delimit a concept. One may, for example, consider the concept of being the brother of Y. This can be explained in terms of the concepts of being male and of being a sibling of Y. These concepts together lay down necessary and sufficient conditions for anything's being a brother. One who grasps these conditions understands precisely what it is to be a brother.

The Republic begins with the use of the dialectical method to discover what justice is. Cephalus proposes the thesis that "justice" means the same as "honesty in word and deed." Socrates searches for and finds a counterexample to this proposal. It is just, he points out, under some conditions, not to tell the truth or to repay debts. If one had borrowed a weapon from an insane person, who then demanded it back in order to kill an innocent person, it would be just to lie to him, stating that one no longer had the weapon. Therefore, "justice" cannot mean the same as "honesty in word" (*i.e.,* telling the truth). By this technique of proposing one definition after another and subjecting each to possible counterexamples, Socrates attempts to find a definition that would be immune to counterexamples. To find such a definition would be to define the concept of justice, and in this way to discover the true nature of justice. In such a case one would be apprehending a form, the common feature that all just things share.

Plato's search for definitions and thereby the nature of forms is a search for knowledge. But how should knowledge in general be defined? In *Theaetetus* Plato argues that it involves true belief. No one can know what is false. A person may mistakenly believe that he knows something, which is in fact false, but this is only thinking that one knows, not knowing. Thus, a person may confidently assert, "I know that Columbus was the first European to land in North America" and be unaware that other Europeans, including Erik the Red, preceded Columbus. So knowledge is at least true belief, but it must also be something more. Suppose that someone believes there will be an earthquake in September because of a dream he had in April and that there in fact is an earthquake in September, although there is no connection between the dream and the earthquake. That person has a true belief about the earthquake but not knowledge. What the person lacks is a good reason supporting his true belief. In a word, the person lacks justification for it. Thus, in *Theaetetus,* Plato concludes that knowledge is justified true belief.

Although it is difficult to explain what justification is, most philosophers accepted the Platonic analysis of knowledge as fundamentally correct until 1963, when the American philosopher Edmund L. Gettier produced a counterexample that shook the foundations of epistemology: suppose that Kathy knows Oscar very well and that Oscar is behind her, out of sight, walking across the mall. Further, suppose that in front of her she sees walking toward her someone who looks exactly like Oscar; unbe-

knownst to her, it is Oscar's twin brother. Kathy forms the belief that Oscar is walking across the mall. Her belief is true, because he is walking across the mall (though she does not see him doing it). And her true belief seems to be justified, because she formed it on the same basis she would have if she had actually seen Oscar walking across the mall. Nonetheless, Kathy does not know that Oscar is walking across the mall, because the justification for her true belief is not the right kind. What her true belief lacks is an appropriate causal connection to its object.

Aristotle. In *Posterior Analytics,* Aristotle (384–22 BC) analyzes scientific knowledge in terms of necessary propositions that express causal relations. Such knowledge takes the form of categorical syllogisms, in which the middle term causally and necessarily connects the major and minor terms. For example, because all stars are distant and all distant objects twinkle, it follows that all stars twinkle. That is, the middle term, "distant objects," connects the minor term, "stars," to the major term, "twinkle," in order to yield the conclusion that all stars twinkle. Aristotle, however, recognizes that not all knowledge is provable. Thus, the premises of the most basic syllogisms are known but not provable. In contrast with scientific knowledge, there is opinion, which is not provable and is about what happens to be true but need not be.

Since the knowledge formulated in syllogisms resides in the mind, which is part of or one faculty of the soul, much of what Aristotle says about knowledge is part of his doctrine about the nature of soul and, in particular, human soul. As he uses the term, every living thing, including plant life, has a soul (*psyche*), a soul being what makes a thing alive. Thus it is important not to equate soul with mind or intellect. The intellect (*nous*) might variously be described as a power, faculty, part, or aspect of the human soul. It should be stressed that for Aristotle the terms soul (*psyche*) and intellect (*nous*) and its constituents were understood to be scientific terms.

Knowledge is something that a person has. Thus it must be in him somewhere, and the location must be his mind or intellect. Yet there can be no knowledge if the knower and the thing known are wholly separate. What then is the relation between the knowledge in the person or his mind and the object of his knowledge? Aristotle's answer is one of his most enigmatic claims. He says, "Actual knowledge is identical with its object."

Here is one suggestion about what Aristotle means. When a person learns something, he acquires something. What he acquires must either be something different from the thing he knows or identical with it. If it is something different, then there is a discrepancy between what he has in mind and the intended object of his knowledge. But such a discrepancy seems to be incompatible with the existence of knowledge. For knowledge, which must be true and accurate, cannot deviate from its object in any way. One cannot know that blue is a colour if the object of that knowledge is something other than that blue is a colour. This idea that knowledge is identical with its object is dimly reflected in the repetition of the variable p in the standard formula about knowledge: S knows that p just in case it is true that p. Although the line of thinking being attributed to Aristotle is defective in several ways, something like it seems to have motivated Aristotle and many other thinkers over the centuries.

To assert that knowledge and its object must be identical raises a question: In what way is knowledge in a person? Suppose that Smith knows Fido. Then Fido is in Smith. Obviously, Fido is not there as he exists in the nonmental world of space and time. In what sense can it be true that a person who knows what a dog is has that object in his mind? Aristotle derives his answer from his general theory of reality. According to him, all (terrestrial) substances

Matter and form

are composed of two principles: matter and form. If there are four dogs—Bowser, Fido, Spot, and Spuds—they are the same in some respect and different in some respect. They are the same in that each belongs to the same kind and each functions similarly. Thus, Aristotle reasons, just as Plato had, that there must be something in virtue of which they are the same, and this he calls "form." That is, Bowser, Fido, Spot, and Spuds each have the very same form of being a dog. They are different in that they are made out of different matter, different parcels of stuff. The form that a thing has is more important than its matter because it is the form that makes the thing what it is. If Fido were to lose the form of being a dog and acquire another, he would no longer be the same thing. The stuff out of which Fido is made is not similarly important, and in fact that stuff changes periodically, as body cells change through metabolic processes, without Fido ceasing to be Fido.

To return to the explanation of knowledge, what is in the knower when he knows what dogs are is the form of being a dog minus the matter. According to Aristotle, matter is literally unintelligible and not essential to what Fido or any other dog is; thus its absence is inconsequential for knowledge, though not for Fido.

In his sketchy account of the process of thinking in *De anima* (*On the Soul*), Aristotle says that the intellect, like everything else, must have two parts: something analogous to matter and something analogous to form. The first of these is the passive intellect; the second is active intellect, of which Aristotle speaks tersely. "Intellect in this sense is separable, impassible, unmixed, since it is in its essential nature activity. . . . When intellect is set free from its present conditions it appears as just what it is and nothing more: it alone is immortal and eternal . . . and without it nothing thinks."

This part of Aristotle's views about knowledge is an extension of what he says about sensation. According to Aristotle, sensation occurs when the sense organ is stimulated by the sense object, typically through some medium, such as light for vision and air for hearing. This stimulation causes a "sensible species" to be generated in the sense organ itself. This "species" is some sort of representation of the object sensed. As Aristotle describes the process, the sense receives "the form of sensible objects without the matter, just as the wax receives the impression of the signet-ring without the iron or the gold." But, since there are different species for each of the five external senses that Aristotle recognized—sight, hearing, touch, taste, and smell—"species" does not mean "image."

Ancient Skepticism. After the development of Aristotle's psychology the next significant event for the theory of knowledge was the rise of Skepticism, of which there were at least two kinds. The first, Academic Skepticism, arose in the Academy after Plato's death and was propounded by the Greek philosopher Arcesilaus (*c.* 315–*c.* 240 BC), about whom the philosophers Cicero, Sextus Empiricus, and Diogenes Laërtius provide information. Academic Skepticism is also called "dogmatic Skepticism" when it is interpreted as arguing for the thesis that nothing is known. The thesis was inspired by Socrates' avowal that the only thing he knew was that he knew nothing. Thus, it asserts that knowledge is impossible. This form of Skepticism seems to be susceptible to an objection raised by the Stoic Antipater (fl. *c.* 135 BC) and others that the view is self-contradictory. To know that knowledge is impossible is to know something; hence, dogmatic Skepticism is false.

Carneades (*c.* 213–129 BC), a member of the Academy, gave a subtle reply. Academic Skepticism, he claimed, should not be interpreted as a claim about how the world is in itself or about a correspondence between thought (or language) and the world, but as a judicial decision. Just as a defendant in a trial does not prove his innocence but relies upon its presumption and defends it against attack, so the Skeptic does not try to prove that he knows nothing but presumes it and defends this presumption against attacks.

Carneades' construal of Academic Skepticism brings it close to the other kind, Pyrrhonism, named after Pyrrho of Elis (*c.* 365–275 BC). None of his works survive, and scholars rely principally on the early 3rd-century-AD writings of Sextus Empiricus to understand Pyrrhonism. Pyrrhonists assert or deny nothing but lead people to give up making any claims to knowledge. The Pyrrhonist's strategy is to show that, for each proposition with some evidence for it, an opposed proposition has equally good evidence supporting it. These arguments for refuting each side of an issue are called "tropes." For example, the judgment that a

Pyrrhonism

tower is round when seen at a distance is contradicted by the judgment that the tower is square when seen up close. The judgment that Providence cares for all things, based upon the orderliness of the heavenly bodies, is opposed by the judgment that many good people suffer misery and many bad people enjoy happiness. The judgment that apples have many properties—shape, colour, taste, and aroma—each of which affects a sense organ, is opposed by the equally good possibility that apples have only one property that affects each sense organ differently.

Pyrrhonists diagnose dogmatism as the unjustifiable preference for one mode of existence over another. Dogmatists prefer wakefulness and sanity over sleep and insanity. But why should sleep and insanity not be the norm? If the dogmatist answers that it is because sleep and insanity involve some deficiency or abnormal physical states, the Skeptic replies, "By what nonquestion-begging criterion are these things said to be deficient or abnormal? Why should insanity not be taken as the primary notion and sanity be defined as the lack of insanity? If it were, then it would not be difficult to see sanity as a deficiency or abnormality, just as insanity currently is. Or why should wakefulness not be seen as the deficient condition in which people do not dream?" The Skeptic does not advocate insanity or sleep but merely argues that a preference for them is no less justified than a preference for sanity and wakefulness.

What is at stake in the preceding Skeptical arguments is "the problem of the criterion," that is, the problem of deciding how one can determine a justifiable standard against which to measure judgments. Truth seems to need a criterion. But every criterion is either groundless or inconclusive. Suppose that something is proffered as a criterion. The Skeptic will ask what proof there is for it. If no proof is offered, the criterion is groundless. If, on the other hand, a proof is produced, a vicious circle begins to close around the dogmatist: What judgment justifies belief in the proof? If there is no judgment, the proof is unsupported; and if there is a judgment, it requires a criterion, which is just what the dogmatist was supposed to have provided in the first place.

If the Skeptic needed to make judgments in order to survive, he would be in trouble. In fact there is another method of survival that bypasses judgment. The Skeptic can live quite nicely, according to Sextus, by following custom and the way things appear to him. In doing this, the Skeptic does not judge the correctness of anything but merely accepts appearances for what they are.

Ancient Pyrrhonism is not strictly an epistemology since it has no theory of knowledge and is content to undermine the dogmatic epistemologies of others, especially of the Stoics and Epicureans. Pyrrho himself was said to have had moral and ethical motives for attacking dogmatists. Being reconciled to not knowing anything, Pyrrho thought, induced serenity (*ataraxia*).

St. Augustine. St. Augustine of Hippo (354–430) claimed that human knowledge would be impossible if God did not illumine the human mind and thereby allow it to see, grasp, or understand ideas. There are two components to his theory: ideas and illumination. Ideas as Augustine construed them are the same as Plato's; they are timeless, immutable, and accessible only to the mind, not to the senses. They are indeed in some mysterious way part of God and seen in God. Illumination, the other element of the theory, was for Augustine and his many followers, at least through the 14th century, a technical term, built upon a metaphor. Since the mind is immaterial, it cannot be literally lighted. Yet the entire theory of illumination rested upon the extended visual metaphor, inherited from Plotinus (205–270) and other Neoplatonic sources, of the human mind as an eye that can see when and only when God, the source of light, illumines it. Still, it is a powerful metaphor relied upon even in the 17th century by René Descartes (*Discourse on Method;* 1637). Varying his metaphor, Augustine sometimes says that the human mind participates in God and even, as in *On the Teacher,* that Christ illumines the mind by dwelling in it. It is important to emphasize that Augustine's theory of illumination concerns all knowledge, and not specifically mystical or spiritual knowledge. In addition to its historical significance, his theory is interesting for showing how diverse epistemological theories have been.

illumination

Before he articulated this theory in his mature years and soon after his conversion to Christianity, Augustine was concerned to refute the Skepticism of the Academy. In *Against the Academicians* Augustine claims that, if nothing else, humans know such disjunctive tautologies as that either there is one world or there is not one world and that either the world is finite or it is infinite. Humans also know many propositions that begin with the phrase "It appears to me that," such as "It appears to me that what I perceive is made up of earth and sky, or what appears to be earth and sky." And they know logical (or what he calls "dialectical") propositions, for example, "If there are four elements in the world, there are not five; if there is one sun, there are not two; one and the same soul cannot die and still be immortal; and man cannot at the same time be happy and unhappy."

Many other refutations of Skepticism occur in later works, notably, in *On the Free Choice of the Will, On the Trinity,* and *The City of God.* In the latter work Augustine proposes other examples of things about which people are absolutely certain. Again in explicit refutation of the Skeptics of the Academy, Augustine argues that if a person is deceived, then it is certain that he exists. Like Descartes, Augustine puts the point in the first person, "If I am deceived, then I exist" (*Si fallor, sum*). A variation on this line of reasoning occurs in *On the Trinity,* when he says that if he is deceived, he is at least certain that he is alive.

Augustine also points out that, since he knows, he knows that he knows; and he notes that this can be reiterated an infinite number of times: If I know that I know that I am alive, then I know that I know that I know that I am alive. This point was codified in 20th-century epistemic logic as the axiom "If X knows that *p,* then X knows that X knows that *p.*" In *The City of God* Augustine claims that he knows that he loves: "For neither am I deceived in this, that I love, since in those things which I love I am not deceived." With Skepticism thus refuted, Augustine simply denies that he has ever been able to doubt what he had learned through his sensations or even the testimony of most people.

Skepticism did not recover from Augustine's criticisms for a thousand years; but then it arose again like the phoenix in Egyptian mythology. Augustine's Platonic epistemology dominated the Middle Ages until the mid-13th century, when St. Albertus Magnus (1200–80) and then his student St. Thomas Aquinas developed an alternative to Augustinian illuminationism.

MEDIEVAL PHILOSOPHY

St. Anselm of Canterbury. The phrase St. Anselm of Canterbury (*c.* 1033–1109) used to describe his own project, namely, "faith seeking reason" (*fides quaerens intellectum*), well characterizes medieval philosophy as a whole. All the great medieval philosophers, Christian, Jewish, and Islāmic alike, were also theologians. Virtually every object of interest was related to their belief in God, and virtually every solution to every problem, including the problem of knowledge, contained God as an essential part. Anselm himself said that, while true propositions are those that signify what is, ultimately truth is God. This presented Anselm with a problem, which he discusses at the beginning of *Proslogium* as a prelude to his famous ontological argument for the existence of God. There is a tension between the view that God is truth and intelligibility and the fact that humans have no perception of God. How can there be knowledge of God, he asks, when all knowledge comes through the senses and God, being immaterial, cannot be sensed? His solution is to distinguish between knowing something by being acquainted with it in sensation and knowing something by describing it. Knowledge by description is possible because of the concepts that one forms from sensation. All knowledge about God depends upon the description that he is "the thing than which a greater cannot be conceived." From this premise Anselm argues that humans can know, for example, that God exists, is all-powerful, all-knowing, all-just, all-merciful, and immaterial. Eight hundred years later Bertrand Russell

would use the same distinction between knowledge by acquaintance and knowledge by description to develop his influential philosophy, although he would have vigorously denied that the distinction could be employed as Anselm had, namely, to prove that God exists.

St. Thomas Aquinas. While a Platonic and Augustinian epistemology dominated the early Middle Ages, the translation of Aristotle's *On the Soul* in the early 13th century had a dramatic effect. Following Aristotle, Thomas Aquinas (1225–74) recognized that there are different kinds of knowledge. Sense knowledge is what results from sensing individual things: thus, one sees a tree, hears the song of an oriole, and tastes or smells a peach. Thomas considered sense knowledge to be low-grade because it has individual things as its object and is also shared with brute animals. Sensation itself does not involve the intellect and is not properly speaking knowledge (*scientia*).

It is characteristic of scientific knowledge to be universal; the more general in scope a piece of knowledge is, the better. This is not to diminish the importance of specificity. Scientific knowledge should also be rich in detail, and God's knowledge is the most detailed. The detail, however, must be essential to the thing being studied and not peculiar to just some instances of that kind. Although Thomas thought that the highest knowledge humans can possess is knowledge of God, knowledge of physical objects is more attuned to human capabilities, and only that kind of knowledge will be discussed here.

In his discussion of knowledge in *Summa theologiae,* Thomas Aquinas argues that human beings do not know material objects directly, nor are such things the principal object of knowledge. Knowledge aims at what is universal, while material things are individual and can be known only indirectly. Elaborating on the thought of Aristotle, Thomas claims that the process of thinking that accompanies knowledge consists of the active intellect (*intellectus agens*) abstracting (*abstrahens*) a concept from an image (*phantasma*) received from the senses.

In one of Thomas' accounts of the process, abstraction is the process of isolating the universal elements of an image of a particular object from those elements that are peculiar to the object. For example, from the image of a dog the intellect abstracts the ideas of being alive, being capable of reproduction, movement, and whatever else might be essential to being a dog. All these ideas are common to all dogs because they are essential to them. These ideas can be contrasted with the ideas of being owned by Dion and weighing five pounds, namely, with properties that vary from dog to dog.

As stated earlier, Aristotle typically spoke of a form as being in the intellect of the knower, whereas the matter of an object is unintelligible and remains extramental. While it was necessary for Aristotle to say something like this in order to escape the absurdity of holding that a material object is in the mind in exactly the same way it is in the physical world, there is also something unsatisfying about it. Physical things contain matter as an essential element, and, if their matter is no part of what is known, then it seems that human knowledge is lacking. In order to counter this worry, Thomas revised Aristotle's theory. He said that not the form alone but the species of an object is also in the intellect. A species is a combination of form and "common matter" (*materia communis*), where common matter is contrasted with individuated matter (*materia signata vel individualis*), which actually gives bulk to a material object. Common matter is something like a general idea of matter. Since every animal must have a body, it is not enough to conceive of an animal merely as something that is alive. Having flesh and bones, that is, being material, is part of the essence of being an animal. Of course this materiality, which is common to every animal, is not the same as the actual flesh and bone that constitute Fido—hence the distinction between common and individuated matter.

Species

This abstracted species resides in a part of the soul called "the passive intellect," where it is described as being illumined by the active intellect. What this process amounts to is the isolation of those features of the intelligible species that are universal and necessary to it. Thus, to know what a human being is is to have abstracted the ideas of being rational and being capable of sensation, movement, reproduction, and nutrition and to have excluded the ideas of living in a particular place or having a certain appearance, all of which are not essential to being human.

One objection that Thomas anticipated being raised against his theory is that it gives the impression that ideas, not things, are what are known. If knowledge is something that humans have and if what humans have in their intellect is a species of a thing, then it is the species that is known and not the thing. It might seem, then, that Thomas' view is a type of idealism.

Thomas had prepared for this kind of objection in several ways. His insistence that what the knower has in his intellect is a species, which includes matter, is supposed to make what is in the intellect seem more like the object of knowledge than an immaterial Aristotelian form. Also, scientific knowledge does not aim at knowing any individual object but at what is common to all things of a certain sort. In this, Thomas' views are similar to those of 20th-century science. The billiard ball that John Jones drops from his porch is of no direct concern to physics. Even though its laws apply to John Jones's ball, physics is interested in what happens to any object dropped from any height, just as what Thomas says about apples in general also applies to each individual apple.

As assuaging as these considerations might be, they do not blunt the main force of the objection. For this purpose Thomas Aquinas introduced the distinction between what is known and that by which it is known. To specify what is known, say, an individual dog, is to specify the object of knowledge; to specify that by which it is known, say, the phantasm or the species of a dog, is to specify the apparatus of knowledge. The species of something is that by which the thing is known; but it is not itself the object of that knowledge, although it can become an object of knowledge by being reflected upon.

The philosophical optimism of the 13th century dissolved as a consequence of the secular and ecclesiastical condemnations in 1270 and 1277 of certain aspects of Aristotelian philosophy, and worries about Skeptical consequences began to emerge. While the philosophy of Thomas Aquinas was one of the targets of these condemnations, John Duns Scotus was also worried about the Skeptical consequences that could be elicited from the major competitor to Aristotelianism, the Augustinianism of Henry of Ghent (1217–93). According to Henry, God must "illumine" the human intellect on every occasion of its knowing. Not only could no good literal meaning be given to this sense of illumination, but the view also sounds as if all human knowledge were supernatural. Henry's insistence that God's illumination is a natural divine illumination did not persuade many people.

John Duns Scotus. While he accepted some aspects of Aristotelian abstractionism and also held that there need to be some a priori principles of perception, principles that he attributed to Augustine, John Duns Scotus (*c.* 1266–1308) did not rest the certainty of human knowledge on either of them. He distinguished four different classes of things that are certain: First, there are things that are knowable simply (*simpliciter*). These include both true identity statements such as "Cicero is Tully" and propositions, later to be called analytic, such as "Man is rational." According to Duns Scotus, such truths coincide with what makes them true. A consequence of this is that the negation of a simple truth is inconsistent even though it may not be explicitly contradictory. For example, the negation of "The whole is greater than any proper part" is not explicitly contradictory in the way that "Snow is white and snow is not white" is; nonetheless, "The whole is not greater than any proper part" cannot possibly be true and hence is contradictory.

The second class of certainly known propositions consists of things knowable through experience, where "experience" has the Aristotelian sense of something that is encountered numerous times. The knowledge afforded by experience is grounded in the a priori epistemic principle that "whatever occurs in a great many instances by a cause that is not free is the natural effect of that cause." It is

important to note that Duns Scotus' pre-Humean confidence in induction did not survive the Middle Ages. The 14th-century philosopher Nicholas of Autrecourt, who has been called "the medieval Hume," argued at length that there is no necessary connection between any two events and that there is no rational justification for the belief in causal relations.

The third class of certainly known propositions consists of things knowable that concern one's own actions (*de actibus nostris*). Humans know when they are awake immediately and not through any inference; they know with certainty that they think (*me intelligere*) and that they hear and have other sense experiences. Even if a sense experience is caused by a defective sense organ, it remains true that one is aware of the sensuous content of the sensation: for example, one sees white even if one is mistaken in thinking that the seeing is caused by snow.

The fourth class of certainly known objects consists of things knowable through human senses (*per sensus*). Duns Scotus said that humans learn about the heavens, the earth, the sea, and all that are in them. This last class of objects that are certainly known things seems to be posited without regard to the threat of Skepticism at all.

Duns Scotus' rendition of intuitive knowledge, however, has the purpose of forestalling the Skeptical move of interposing something between the knower and the thing known that might enable belief to deviate from its object. Intuitive knowledge is indubitable knowledge that something exists. It is knowledge "precisely of a present object [known] as being present and of an existent object [known] as being existent." Further, the object of knowledge must be the cause of the knowledge. If a person sees Socrates before him, then, according to Duns Scotus, he has intuitive knowledge of the proposition that Socrates is white and that Socrates and his whiteness cause that knowledge. Intuitive knowledge contrasts with abstractive knowledge, such as knowledge of universals, for which the object need not be present or even existent. For example, for all one knows from contemplating the nature of dogs or unicorns, they are equally likely or unlikely to exist.

ıtuitive
nowledge

It may appear that intuitive knowledge is absolute; either one has it or one does not. But that is not Duns Scotus' doctrine. He held that there is imperfect intuitive knowledge of the past, which is more certain than abstractive knowledge but less certain than present intuitive knowledge. However plausible or implausible this may be, it is worth noting that Russell held the same view but expressed it by using the terms "knowledge by acquaintance" (intuitive cognition) and "knowledge by description" (abstractive cognition).

William of Ockham. There are several places in Duns Scotus' account where Skeptical challenges can gain a foothold, for example, when he endorses the certainty of sense knowledge and when he holds that intuitive cognition must be of an existent object. William of Ockham (*c.* 1285–1349?) took his stand against the Skeptical challenge by radically revising Duns Scotus' idea of intuitive cognition. Unlike Duns Scotus, Ockham does not require intuitive knowledge to have an existent object, and the object of intuitive knowledge need not be its cause. To the question "What is the basis for the distinction between intuitive and abstractive knowledge?" given that it is not the existence of the object and not a causal relation between an object and the knower, Ockham answered that they are simply different. His answer notwithstanding, it is characteristic of intuitive knowledge that it is unmediated. There is no gap between the knower and the known that might undermine certainty: "I say that the thing itself is known immediately without any medium between itself and the act by which it is seen or apprehended."

atural
ıd
ıper-
ıtural
tuitive
ıowledge

According to Ockham, there are two kinds of intuitive knowledge: natural and supernatural. In natural intuitive knowledge, the object exists, the knower judges that the object exists, and the object causes the knowledge. In supernatural intuitive knowledge, the object does not exist, the knower judges that the object does not exist, and God is the cause of the knowledge. In neither case is knowledge a relation; it is something a person has, a property of the person.

Ockham recognized that God might cause a person to think that he has intuitive knowledge of an existent object when there in fact is no such object. But such a condition is not intuitive knowledge but a false belief. Unfortunately, in acknowledging that a person has no way to distinguish between genuine intuitive cognitions and divine counterfeits of them, Ockham has in effect lost the argument to the Skeptics.

Later medieval philosophy followed a fairly straight path to Skepticism. John of Mirecourt was condemned in 1347 for holding among other things that there is no certainty of external reality because God could cause illusions to seem real. Nicholas of Autrecourt was also condemned in the same year for holding that only purely sensory reports of human experience are certain and that the only certain principle is that of contradiction, namely, that a thing cannot be and not be something at the same time. He denied that humans know that causal relations exist or that there are substances, two of many errors he credited to Aristotle, about whom he said, "In all his natural philosophy and metaphysics, Aristotle had hardly reached two evidently certain conclusions, perhaps not even a single one. . . ." The link between Skepticism and criticism of Aristotle was fairly strong, and Petrarch, in *On My Ignorance and That of Many Others* (1367), cited Aristotle as "the most famous" of those who do not have knowledge.

From scientific theology to secular science. For most of the Middle Ages there was no split between theology and science (*scientia*). Science was knowledge that was deduced from self-evident principles, and theology received its principles from the source of all principles, God. In every way, theology was superior to the other sciences, according to Thomas Aquinas. By the 14th century the ideas of science and theology began to be separated. Roughly, theologians began to argue that human knowledge was much more narrowly circumscribed than earlier believed. They often exploited the omnipotence of God in order to undercut the arrogance and pretension of human reason. Their motive was to enhance the dignity of God at the expense of human reason, and in place of rationalism in theology, they promoted a kind of fideism.

Gregory of Rimini (*c.* 1300–*c.* 1358) exemplified the growing split between natural reason and theology. According to Gregory, theology is not a science, and theological propositions are not scientific. In the new view of Gregory, who was inspired by Ockham, science deals only with what is accessible to humans through natural means, that is, through the ordinary operations of their senses and intelligence. Theology in contrast deals with what is accessible in some supernatural way. Thus, theology is not scientific. The role of theology is to explain the meaning of the Bible and the articles of faith and to deduce conclusions from them. Since the credibility of the Bible rests upon belief in divine revelation and revelation upon the authority of God, theology lacks a rational foundation. Further, since there is neither self-evident knowledge of God nor any natural experience of him, humans can have only an abstract understanding of what he is.

Ockham and Gregory did not at all intend their views to undermine theology. For them, natural science is built on probabilities, not certainties. Since humans are fallible, their natural science is fallible, unlike theology, which is built upon propositions that have the authority of God. Unfortunately for theology, the prestige of natural science rose in the 16th century and skyrocketed in the 17th and 18th centuries; modern thinkers preferred coming to their own conclusions based upon experience and reason, even if these were only probable, to trusting the authority of anyone, even God. (This attitude has been called "the Faustian ethos," after Goethe's character Faust.) As the theologians tended to lose confidence in reason, other thinkers who had no or virtually no commitment to Aristotelian thought became the champions of reason and helped give birth to modern science.

MODERN PHILOSOPHY

Faith and reason. Modern philosophers as a group are usually thought to be purely secular thinkers. Nothing could be further from the truth. From the early 17th cen-

tury until the middle of the 18th century, all of the great philosophers incorporated substantial religious elements into their work. Descartes, in his *Meditations* (1641), offered two different proofs for the existence of God, and he asserted that no one who does not believe in a cogent proof for the existence of God can have knowledge in the proper sense of the term. Benedict Spinoza began his *Ethics* (1677) with a proof for the existence of God, after which he expatiated on its implications for understanding all reality. And George Berkeley explained the stability of the sensible world by relying upon God's constant thought of it.

Among the reasons modern philosophers are mistakenly thought to be primarily secular thinkers is that many of their epistemological principles, including some that were intended to defend religion, were later interpreted as subverting the rationality of religious belief. The role of Thomas Hobbes (1588–1679) and John Locke might be briefly considered in this connection. In contrast with the standard view of the Middle Ages that propositions of faith are rational, Hobbes argued that propositions of faith belong not to the intellect but to the will. To profess religious propositions is a matter of obeying the commands of a lawful authority. One need not even understand the meanings of the words professed: an obedient mouthing of the appropriate confession of faith is sufficient. In any case, the linguistic function of virtually every religious proposition is not cognitive in the sense of expressing something that is intended to represent a fact about the world but rather to give praise and honour to God. Further, in contrast to the medieval view, according to which theology is the highest science, theology is not a science at all since its propositions are not susceptible to rational dispute.

The subordination of religious propositions to reason

In *An Essay Concerning Human Understanding* (1689), Locke further eroded the intellectual status of religious propositions by making them subordinate to reason in several dimensions. First, reason can dictate what the possible content of a proposition allegedly revealed by God might be; in particular, no proposition of faith can be a contradiction. Consequently, if the proposition that Jesus is both fully God and fully man is contradictory, it cannot be revealed and cannot be a matter of faith. Also, no revelation can be communicated that contains an idea not based upon sense experience. Thus, St. Paul's experience of things "as eye hath not seen, nor ear heard, nor hath it entered into the heart of man to conceive," are things in which other people can have no faith. To move to another dimension in which reason takes precedence over faith, direct sense knowledge (what Locke calls "intuitive knowledge") is always more certain than any alleged revelation. Thus, a person who sees that someone is soaking wet cannot have it revealed to him that the person is at that moment dry. Rational proofs, in mathematics and science, also cannot be contradicted by divine revelation. The interior angles of a rectangle equal 360°, and no alleged revelation to the contrary is credible. In short, "*Nothing that is contrary to, and inconsistent with, the clear and self-evident dictates of reason, has a right to be urged or assented to as a matter of faith. . . .* "

What space, then, does faith occupy within the mansion of human beliefs? According to Locke, it shares a room with probable truths, those propositions of which reason cannot be certain. There are two types: claims about observable matters of fact and claims that go "beyond the discovery of our sense." Religious propositions belong to each category, as do empirical or scientific ones. That Caesar crossed the Rubicon and that Jesus walked on water belong to the first type of probable proposition. That heat is caused by the friction of imperceptibly small bodies and that angels exist are propositions that belong to the second category.

While mixing religious claims with scientific ones might seem to secure a place for the former, in fact it did not. For Locke also held that whether something is a revelation or not "reason must judge," and more generally that "*Reason must be our last judge and guide in everything.*" Although this maxim was intended to reconcile reason and revelation—indeed, he calls reason "natural revelation" and revelation "*natural reason enlarged* by a new set of discoveries communicated by God"—over the course of 200 years reason repeatedly judged that alleged revelations had no scientific or intellectual standing.

Although there is a strong religious element in modern thinkers, especially before the middle of the 18th century, the purely secular aspects of their thought predominate in the following discussion, because it is these that are of contemporary interest to epistemologists.

Impact of modern science on epistemology. Nicolaus Copernicus (1473–1543), a cleric, argued in *On the Revolutions of the Celestial Spheres* (1543) that the Earth revolves around the Sun. His theory was epistemologically shocking for at least two reasons. First, it goes directly counter to how humans experience their relation to the Sun; it is everyone's prescientific view that the Sun revolves around the Earth. If science can overthrow such a belief, then scientific reasoning seems to lead to knowledge in a way that nonscientific reasoning cannot. Indeed, the nonscientific reasoning of everyday life may seem to be a kind of superstition. Second, his theory was shocking because it contradicts the view that is presented in several books of the Bible, most importantly the explicit account in Genesis of the structure of the cosmos, according to which Earth is at the centre of creation and the Sun hangs from a celestial ceiling that holds back the waters which once flooded the Earth. If Copernicus is right, then the Bible can no longer be taken as a reliable scientific treatise. Scientific beliefs about the world, then, must be gathered in a radically new way.

Many of the discoveries of Galileo Galilei (1564–1642) had the same two shocking consequences. His telescope seemed to reveal that unaided human vision gives false or seriously incomplete information about the nature of celestial bodies. His mathematical formulations of physical phenomena seem to indicate that most sensory information may contribute nothing to knowledge. Like his contemporary, the astronomer Johannes Kepler, he distinguished between two kinds of properties. Primary qualities, such as shape, quantity, and motion, are genuine properties of things and are knowable by mathematics. Secondary qualities, namely, odour, taste, sound, colour, warmth, or coldness, exist only in human consciousness and are not part of the objects to which they are normally attributed.

René Descartes. Both the rise of modern science and the rediscovery of Skepticism were important influences on René Descartes (1596–1650). While he believed that humans were capable of knowledge and certainty and that modern science was developing the superstructure of knowledge, he thought that Skepticism presented a legitimate challenge that needed an answer, one that only he could provide.

The challenge of Skepticism, as Descartes saw it, is vividly portrayed in his *Meditations.* He considered the supposition that all of one's beliefs are false, being the delusions of an evil genius who has the power to impose beliefs on people unbeknownst to them. But Descartes claimed that it is not possible for all of one's beliefs to be false, for anyone who has false beliefs is thinking and knows that he is thinking, and if the person is thinking, then that person exists. Nonexistent things cannot think. This line of argument is summarized in Descartes's formula, "Cogito, ergo sum" ("I think; therefore, I am").

Descartes' refutation of Skepticism

Descartes distinguished two sources of knowledge: intuition and deduction. Intuition is an unmediated mental seeing or direct apprehension of something experienced. The truth of the proposition "I think" is guaranteed by the intuition one has of one's own experience of thinking. One might think that the proposition "I am" is guaranteed by deduction, as is suggested by the "ergo." In *Objections and Replies* (1642), however, Descartes explicitly says that the certainty of "I am" is also based upon intuition.

If one could know only that one thinks and exists, human knowledge would be depressingly narrow. So Descartes proceeded to broaden the limits of human knowledge. After showing that all human knowledge depended upon thought or reason, not sensation or imagination, he then proceeded to prove to his own satisfaction that God exists;

that the criterion for knowledge is clearness and distinctness; that mind is more easily known than body; that the essence of matter is extension; and that most of his former beliefs are true.

Few of these proofs convinced many people in the form in which Descartes presented them. One major problem is what has come to be known as the Cartesian circle. In order to escape from the possibility that an evil genius is deluding him about everything he believes, Descartes proves that God exists. He then argues that clearness and distinctness is the criterion for all knowledge because God does not deceive man. But, since this criterion is arrived at only after the existence of God has been proven, he cannot appeal to this criterion when he presents his proof for the existence of God; hence he cannot know that his proof is cogent.

John Locke. *An Essay Concerning Human Understanding* by John Locke (1632–1704) is often taken to be the first major empiricist work. Book I discusses innate ideas in order to deny that there are any; Book II discusses various genuine kinds of ideas; Book III discusses language with an emphasis on the meaning of words; and Book IV discusses knowledge and related cognitive states and processes.

Innate ideas are ideas that humans are born with. Rationalist philosophers, like Descartes and Gottfried Wilhelm Leibniz (1646–1716), thought that there have to be such ideas in order to explain the existence of some of the ideas which humans have. One argument for innate ideas is that, while the ideas of blue, dog, and large, for example, can be explained as the result of certain sense impressions, other ideas seem unable to be attributed to sensation. Numbers, for example, seem to be outside the realm of sensation. Another argument is that some principles are accepted by all human beings, as, for example, the principle that out of nothing nothing comes. Locke did not think either of these arguments had any force. He held that all ideas can be explained in terms of sensation, and he set as one of his projects the task of providing such an explanation. Instead of directly attacking the hypothesis of innate ideas, Locke's strategy was to refute it by showing that it is otiose and hence dispensable.

In Book II of the *Essay* Locke supposes the mind to be like a blank sheet of paper that is to be filled with writing. How does the paper come to be filled? "To this I answer, in one word," says Locke, "Experience." He divides experience into two types: observation of external objects and observation of the internal operations of the mind.

Observation of external objects is another description for sensation. Observation of the internal operations of the mind does not have its own word in ordinary language, and Locke stipulated "reflection" to designate it, because people arrive at ideas by reflecting on the operations of their own minds. Examples of reflection are perceiving, thinking, doubting, believing, reasoning, knowing, and willing.

An idea is anything that the mind "perceives *in itself,* or is the immediate object of perception." Qualities are the powers that objects have to cause ideas. Many words have dual senses. The word *red,* for example, might mean either the idea of red in the mind or the quality in a body that causes the idea of red in the mind. Some qualities are primary in the sense that all bodies have them. Solidity, extension, figure, and mobility are primary qualities. Secondary qualities are those powers that, in virtue of the primary qualities, cause the sensations of sound, colour, odour, and taste. Locke's view is that the phenomenal redness of a fire engine is not in the fire engine itself, nor is the phenomenal sweet smell of a rose in the flower itself. Rather, certain configurations of the primary qualities cause phenomena such as the appearance of red or the taste of sweetness, and in virtue of these configurations the object itself is said to have the quality of redness or sweetness. But there is no resemblance between the idea in the mind (phenomenon) and the secondary quality that causes it. Locke claims, without justification, as George Berkeley was later to argue, that there is, however, a resemblance between primary qualities and the ideas of them. (Locke distinguishes a third sort of quality, *e.g.,* the power of fire to produce a new colour or consistency in wax or clay, but he makes nothing of it.)

Primary and secondary qualities

Although Locke along with most distinguished modern philosophers repudiated Aristotelianism and the Scholasticism to which it gave rise, a doctrine of abstraction survives in his philosophy. Abstraction occurs when "ideas taken from particular beings become general representatives of all of the same kind." That is, to abstract is to ignore the particular circumstances of time and place and to use an idea to represent all things of a certain kind.

In Book IV Locke finally defines knowledge as "*the perception of the connexion of and agreement, or disagreement and repugnancy of any of our ideas.*" He also distinguishes several degrees of knowledge. The first is knowledge in which the mind "perceives the agreement or disagreement of two ideas *immediately by themselves,* without the intervention of any other," which he calls "intuitive knowledge." His first examples are such analytic propositions as "*white* is not *black,*" "a *circle* is not a *triangle,*" and "*three* are more than *two.*" But later he says, "The knowledge of our own being we have by intuition." Relying on the metaphor of light as Augustine and others had, Locke says of this knowledge that "the mind is presently filled with the clear light of it. *It is on this intuition that depends all the certainty and evidence of all our knowledge.*"

The second degree of knowledge occurs when "the mind perceives the agreement or disagreement of . . . ideas, but not immediately." Some mediating idea makes it possible to see the connection between two other ideas. Proofs are things that show the mediating connections between ideas, and a clear and plain proof is a demonstration. Demonstrative knowledge is certain but not as evident as intuitive knowledge, says Locke, because it requires effort and attention to go through the steps needed to recognize the certainty of the conclusion.

A third degree of knowledge, "sensitive knowledge," approximates to what Duns Scotus and Ockham called "intuitive cognition," namely, the perception of "*the particular existence of finite beings without us.*" Unlike medieval intuitive cognition, Locke's sensitive knowledge is less certain than his intuitive or demonstrative knowledge.

Beneath knowledge is probability, which is the appearance of agreement or disagreement of ideas with each other. Etymologically, probability is a likeliness to be true, and it guides in matters "whereof we have no certainty." Locke suggests that probability rests upon the testimony of others and, like knowledge, comes in degrees, which depend upon the likely veracity of the sources of the proposition. The highest degree of probability attaches to propositions endorsed by the general consent of all people in all ages. Locke may have in mind the virtually general consent of his contemporaries in the proposition that God exists. But he explicitly mentioned beliefs about causal relations, which are not perceived but inferred. To argue from such beliefs is called "an argument from the nature of things." The next degree of probability or assurance in probable propositions attaches to matters that hold not universally but for the most part, such as that persons prefer their own private advantage to the public good. This sort of proposition is typically derived from history. The next degree of probability or assurance attaches to claims about specific facts, for example, that a man named Julius Caesar lived a long time ago. Problems arise when testimonies conflict, as they often do, but there is no simple rule or set of rules that instructs one how to resolve such controversies.

In addition to these probabilities, all of which concern particular matters of fact, there are also probabilities about things that are not within the power of the senses. The existence, nature, and operation of angels, devils, microbes, magnets, and molecules all fall into this class. It is important to recognize that for people as scientific as Locke, who was a member of the Royal Society, all of these were part of the same class. It took many centuries to separate science from religion and superstition.

George Berkeley. Locke is part of a philosophical tradition called empiricism, that is, the view that the sole or at least the major source of human knowledge is sensory experience. George Berkeley (1685–1753) was the next

great adherent of empiricism. In his major work, *Treatise Concerning the Principles of Human Knowledge* (1710), he divides ideas into three types: Ideas that come from sense correspond to Locke's simple ideas of sensation. Ideas that come from "attending to the passions and operations of the mind" correspond to Locke's ideas of reflection. Ideas that come from compounding, dividing, or otherwise representing ideas, correspond to Locke's compound ideas. An apple, for example, is a compound of the simple ideas of colour, taste, smell, and figure associated with it.

In addition to ideas, what exists are spirits or souls or minds. By "spirit," Berkeley means "one simple, undivided, active being." Spirit exercises itself in two ways: in understanding and in willing. Understanding is spirit perceiving ideas, and will is spirit producing ideas. It is evident, says Berkeley, that no idea, including those of sensation, can exist outside of a mind. This is evident, not merely in virtue of the meaning of "idea" but what it means to exist. For a table to exist is for someone to see or feel it. To be an odour is to be smelled. To be a sound is to be heard. In short, for nonthinking beings, *esse* is *percipi* (to be is to be perceived).

The question whether a tree falling in a virgin forest makes a sound is inspired by Berkeley's philosophy, though he never asked it in those terms. He did, however, consider the thrust of the objection and gave various answers to it. He sometimes says that a table in a room unperceived is a table that would be perceived if someone were there. This conditional response, however, is not sufficient. Granted that the table would exist if it were perceived, does it exist when it is not perceived? Berkeley's other answer is that, when no human is perceiving a table or other such object, God is; and it is his thinking that keeps the otherwise unperceived object in existence.

However strange his doctrine may initially sound, Berkeley claimed that he was merely describing the commonsense view of reality. To say that colours, sounds, trees, dogs, and tables are ideas is not to say that they do not really exist. It is merely to say what they are. To say that animals and pieces of furniture are ideas is not to say that they are diaphanous, gossamer, and evanescent. Opacity, density, and permanence are also ideas that partially constitute these objects.

Berkeley has a syllogistic argument for his main point: physical things, such as trees, dogs, and houses, are things perceived by sense, and things perceived by sense are ideas; therefore, physical things are ideas. If one objects that the first premise is false, Berkeley in reply would challenge the objector to point out one example of something that is not sensed. The only way to identify such an example is through some sensation, either by sight, touch, taste, or hearing. In this way, any proffered counterexample becomes an example of Berkeley's point.

If one objects that the second premise of the syllogism is false on the grounds that people sense things, not ideas, Berkeley would reply that there are no sensations without ideas and that it makes no sense to speak of some additional thing which ideas are supposed to represent or resemble. Unlike Locke, Berkeley does not believe that there is anything "behind" ideas in a world external to the mind. There could not be. If the alleged external objects, of which ideas are supposed to be representations, exist, then they are themselves either ideas or not. If they are ideas, then Berkeley's point that everything perceived is an idea is vindicated. If they are not ideas, then they are unperceived; in particular, they would be invisible colours, intangible textured things, odourless smells, and silent sounds. If someone objects that he can imagine trees or books in a closet unperceived, Berkeley would reply that this proves nothing except that there are imagined trees and books. People who think that there are unperceived objects are deceived because they do not take into account their own thinking of the allegedly unperceived object.

A consequence of this argument is that Locke's distinction between primary and secondary qualities is spurious. Extension, figure, motion, rest, and solidity are as much ideas as green, loud, and bitter are; there is nothing special about the former kinds of ideas. Furthermore, matter, as philosophers conceive it, does not exist and indeed is contradictory. For matter is supposedly unsensed extension, figure, and motion, but since extension, figure, and motion are ideas, they must be sensed.

Berkeley's doctrine that things unperceived by human beings continue to exist in the thought of God was also not novel. It was part of the traditional belief of Christian philosophers from Augustine onward through Aquinas and at least to Descartes that God not only creates all things but keeps them in existence by thinking of them. In this view if he were ever to stop thinking of a creature, it would immediately be annihilated.

On another matter, the doctrine of abstraction, Berkeley made a clean break with the past. Berkeley rejected it completely, because he thought it led to belief in unperceived, nonspiritual substances. Abstractionism, according to Berkeley, illicitly warrants the separation of existence from being perceived. For him every idea is particular and of a particular object. There cannot be an idea of motion in general but only of a certain body moving slowly or quickly. To reject abstract ideas is not to reject general ideas. An idea is general in virtue of "being made to represent or stand for all the other particular ideas of the same sort." That is, each general idea is a particular idea that stands for many things.

David Hume. Although Berkeley rejected the Lockean notions of primary and secondary qualities and matter, he retained Locke's beliefs in the existence of mind, substance, and cause as a power or secret force. David Hume (1711–76), in addition to rejecting all the Lockean notions that Berkeley did, also rejected what Berkeley had retained. His justification for this step was empiricist and scientific, for he thought that all science is empiricist and that there is no empirical justification for belief in mind or spirit.

Hume aspired to be the Newton of philosophy. As stated in *A Treatise of Human Nature* (1730–40), he wanted to formulate universal principles to explain "all effects from the simplest and fewest causes," but a boundary condition on these principles is that they "cannot go beyond experience." Further, the ultimate principles that humans can form will themselves lack justification. They will explain experience without having an explanation of their own.

Kinds of perceptions. Hume has a twofold division of perceptions: impressions and ideas. Impressions are perceptions that enter with "most force and violence." Ideas are "faint images" of impressions. Hume thinks the distinction so obvious that he demurs from explaining it at any length. Impressions are felt; ideas are thought, he indicates in his summary explication. He also concedes that, although one can always discern the difference between an impression and an idea by its force, sleep, fever, and madness sometimes produce ideas that approximate to the force of impressions, and certain impressions approach the weakness of ideas. But such occasions are rare.

The distinction has a problem that Hume did not notice. The impression (experience) of anger has an unmistakable quality and intensity, but it is not the case that the idea of anger always makes a person feel angry. Thinking of anger no more guarantees being angry than thinking of the idea of happiness guarantees being happy, even if thinking happy thoughts tends to make people happy. So there is a difference between the experience of anger and the idea of anger that Hume's philosophy does not capture.

In addition to impressions and ideas, perceptions can be divided into the categories of simple and complex. Whereas simple perceptions are not subject to further separation or distinction, complex perceptions are. For example, apples, although unitary objects in one view, are in fact complex perceptions; they are divisible into a certain shape, colour, texture, and aroma. It is noteworthy that for every simple impression there is a simple idea that corresponds to it and differs from it only in force and vivacity, and vice versa. So, corresponding to the impression of red is the idea of red. This does not hold true in general for complex perceptions. Although there is a correspondence between the impression of an apple and the idea of an apple, there is not always a correspondence between impressions and ideas. There is no impression that corresponds to the idea of Pegasus or a unicorn; these complex ideas do not have a correlate in reality. There are also complex impressions

that do not have a corresponding idea. A traveler who has seen an extensive part of Rome nonetheless does not have an idea of Rome that corresponds in every respect to his perceptions.

Because of their correspondence, there seems to be a special connection between simple impressions and simple ideas: the former cause the latter. Hume deduces this on the following grounds. A simple impression always precedes the corresponding idea, and the idea invariably follows the conjoined impression. Thus, because of the temporal priority of impressions and the constant conjunction of impressions and ideas, Hume concludes that impressions cause ideas.

There are two kinds of impressions: sensation and reflection. Sensation "arises in the soul originally from unknown causes." Hume says little more about sensation because discussion of it belongs to anatomists and scientists. (Many late 20th-century philosophers do not accept this division between philosophy and anatomy.) To explain reflection is rather complicated because it derives from a complex mental operation. After people feel heat or cold, thirst or hunger, pleasure or pain, they form ideas of heat or cold, thirst or hunger, pleasure or pain. And, following the formation of these ideas—at a third stage of cogitation—they form from the ideas the second kind of impressions: impressions of "desire and aversion, hope and fear." These impressions are the result of reflecting on ideas caused by sensation.

Reflection

Since imagination can divide and assemble disparate ideas as it will, some explanation is needed for why the mind seems to run in predictable channels. Hume says that the mind is guided by three principles: resemblance, contiguity, and cause and effect. Thus a person who thinks of one idea is likely to think of another idea that resembles it. For example, a person's thought, if one accepts Hume's account, will run from red to pink to white, or from dog to wolf to coyote. Hume also uses the principle of resemblance to explain how general ideas function. Hume agrees with Berkeley in denying that there are abstract ideas, and he affirms that all ideas are particular. Some of them, however, are used to represent many objects by inclining the mind to think of other ideas that resemble the first. These particular ideas that represent many things are general ideas. Concerning contiguity, people are inclined to think of things that are next to each other in space and time. Finally and most importantly, people associate ideas on the basis of cause and effect relations. Fire and smoke, parent and child, disease and death are tied in the mind because of their causal relations. But cause and effect relations play a more central role in Hume's thought than these brief remarks might suggest.

Cause and effect. Although people gain much information from their impressions, most matters of fact depend upon reasoning about causes and effects, even though people do not directly experience causal relations. What, then, are causal relations? According to Hume they have three components: contiguity of time and place, temporal priority of the cause, and constant conjunction. In order for x to be the cause of y, x and y must exist adjacent to each other in space and time, x must precede y, and x and y must invariably exist together. There is nothing more to the idea of causality than this; in particular, people do not experience and do not know of any power, energy, or secret force that causes possess and that they transfer to the effect. Still, all judgments about causes and their effects are based upon experience. To cite examples from *An Enquiry Concerning Human Understanding* (1748), since there is nothing in the experience of seeing a fire close by which logically requires that one will feel heat, and since there is nothing in the experience of seeing one rolling billiard ball contact another that logically requires the second one to begin moving, why does one expect heat to be felt and the second ball to roll? The explanation is custom. In previous experiences, the feeling of heat has regularly accompanied the sight of fire, and the motion of one billiard ball has accompanied the motion of another. Thus the mind becomes accustomed to certain expectations. "All inferences from experience, therefore, are effects of custom, not of reasoning." Thus it is that custom, not reason, is the great guide of life. In short, the idea of cause and effect is neither a relation of ideas nor a matter of fact. Although it is not a perception and not rationally justified, it is crucial to human survival and a central aspect of human cognition.

Substance. One of the cornerstones of philosophy from Plato to Berkeley was the notion of substance, that which exists in itself and does not depend upon anything else for its existence. Substance is contrasted with accident or modes of being, which exist in substances and depend on them for their existence. A dog is a substance, and its colour, shape, weight, and bark exist in the dog and depend on it for their existence. One of the reasons for Hume's place in the history of philosophy is that he denied the existence of substance, using the epistemological principles he shared, not simply with empiricists like Locke and Berkeley, but with Aristotle and Aquinas as well. As argued in the *Treatise,* since all human knowledge must be traced back to sensation, the idea of substance must be also. But what sensation can give rise to the idea of substance? It is not a colour, shape, sound, or taste. Substance, by its proponents' own definition, is not an accident or mode. Hume concludes, "We have therefore no idea of substance, distinct from that of a collection of particular qualities, nor have we any other meaning when we either talk or reason concerning it." What then are the things that earlier philosophers designated substances? They are "nothing but a collection of simple ideas, that are united by the imagination, and have a particular name assigned to them." Gold, to take Hume's example, is nothing but the collection of the ideas of yellow, malleable, fusible, and so on. Even the mind is only a collection, "a heap or collection of different perceptions united together by certain relations and suppos'd tho' falsely, to be endow'd with a perfect simplicity or identity."

Relations of ideas and matters of fact. Human thought concerns two kinds of things: relations of ideas and matters of fact. Relations of ideas can either be intuited, that is, seen directly, or deduced from other propositions. That a is identical with a, that b resembles c, and that d is larger than e are examples of propositions that are intuited. The opposites of true propositions expressing relations of ideas are contradictory. Arithmetic and algebra are the subjects about which there can be the most certainty. In his *Treatise* Hume says that geometry is almost as certain as these, but not quite, because its original principles derive from sensation, and about sensation there can never be absolute certainty. He revised his views about geometry later, and in the *Enquiry* he puts geometry on an equal footing with the other mathematical sciences.

In contrast with relations of ideas, matters of fact are derived from experience. Experience, however, would be quite limited if it did not include causal relations, which go beyond what is experienced.

Skepticism. Hume's discussion about relations of ideas and matters of fact gives the impression that he thought that human knowledge is possible. Relations of ideas seem to be the object of knowledge, while matters of fact seem to be the object of probability. In Part II of the *Treatise* he denies this and argues forcefully for Skepticism.

Until the beginning of Part IV of Book I of the *Treatise,* there is little or no hint of Skepticism. The distinction between knowledge (of the mathematical sciences) and probability (of matters of fact) seems to presuppose that there is knowledge. But one then discovers that Skepticism undermines it all. Although the rules of science are certain and infallible, the application of those rules by humans is uncertain and fallible because humans are prone to error. It does no good for a person to try to check his chain of reasonings because the process of checking is no more immune to error than the original calculation. How can one know that the checking process was performed correctly? And, if the checking procedure seems to identify a mistake in the original calculation, how can one determine whether the error is in the original or in the seeming identification of an error? Adding a checking procedure is in one respect worse than leaving the original calculation alone. It introduces a second event, which, like the original calculation, is possibly flawed. And it is more probable

that one of two possibly flawed events is flawed than either one of the two alone. "By this means," Hume says, "all knowledge degenerates into probability." Another way to see this consequence is to consider that reason is a cause of truth. But, since all causal relations are probable, not certain, all human reasoning is at best probable.

If one thinks further about the matter, the probability of knowledge diminishes and doubt increases. Each judgment of the probability of some judgment introduces further reasons for doubt and thus lowers the overall probability. The joint probability of p and q is lower than the probability of p; and the joint probability of p, q, and r is lower than the probability of p and q. Ultimately, "when I proceed still farther, to turn the scrutiny against every successive estimation I make of my faculties all the rules of logic require a diminution, and at last a total extinction of belief and evidence." If one should say, "Surely, you are kidding," Hume's answer would be a beguiling one: In a sense, "yes," for nature has so made human beings that they cannot in fact be skeptical even though the argument for Skepticism is cogent. As Hume says in his *Enquiry,* people conduct their lives for the most part governed by custom and nature, not reason. Skepticism is true even though there are no Skeptics, because, as in Berkeley's philosophy, the arguments for Skepticism "*admit of no answer and produce no conviction.*"

There is another way of expressing Hume's position. If one examines the grounds that human beings have for trusting their reasoning, one will not be able to find rational grounds. Reason cannot be rationally grounded, and the ground of rationality is wholly nonrational: "*belief is more properly an act of the sensitive, than of the cogitative part of our natures.*"

Some people have tried to make short shrift of Skepticism by pointing out that if the Skeptic recognizes his arguments to be rationally compelling, then he must recognize the sovereignty of reason and hence the falsity of Skepticism. Hume points out that the battle against Skepticism cannot be won in this way. Skepticism is a refutation of the claims of reason. As such, one assumes the truth of rationalism in order to show that it is contradictory and hence false. In other words, Hume's proof is a *reductio ad absurdum* argument against belief in rationality. The Skeptical argument proceeds by arguing that, if rationalism is true, then it is not rational to be rational. Since the consequent is contradictory, the assumption that rationalism is true must be false. Thus, rationalism is false.

Hume has been called "the complete Pyrrhonist," but Hume himself denied that he was one, in large part because he did not distinguish between Pyrrhonism and Academic Skepticism, both of which, according to him, advocate the suspension of belief even as one conducts one's ordinary affairs. Hume thought such a program impossible for human beings: humans are condemned to believe. Unlike the Pyrrhonist, Hume does not suspend judgment or abandon reason. He judges according to reason because it is his nature to do so even though Skeptical arguments against reason are cogent. This philosophical schizophrenia—the use and trust in reason coupled with the recognition that rationality has no rational justification—is part of what Hume calls "mitigated Skepticism." Another part of it is restricting one's investigations to topics that are within the "narrow capacity of human understanding," namely to experience and the mathematical sciences.

Mitigated Skepticism

Given his Skepticism, one might wonder whether it could be directed against Hume's own positive doctrine. It can. At the end of Part I of his *Treatise* Hume says, "Can I be sure, that in leaving all establish'd opinions I am following truth; and by what criterion shall I distinguish her, even if fortune shou'd at last guide me on her foot-steps? After the most accurate and exact of my reasonings, I can give no reason why I shou'd assent to it; and feel nothing but a *strong* propensity to consider objects *strongly* in that view, under which they appear to me." Ultimately one judges according to custom and the way nature dictates one must judge. The *Enquiry Concerning Human Understanding* is intended to be an accurate description of how people judge, not a justification of it.

Immanuel Kant. Idealism is often defined as the view that everything which exists is mental; that is, everything is either a mind or depends for its existence upon a mind, as do ideas and thinking. Immanuel Kant (1724–1804) was not strictly an idealist according to this definition, although he called himself a "transcendental idealist." On his view, humans can know only what is presented to their senses or what is contributed by their own mind. Every sensory experience is a mixture of a sensory content, which is simply given to a person, and a spatial and temporal form, which is contributed by the mind itself. Further, if one formulates a sensory experience into a judgment, then the mind also contributes certain additional objective features: the judgment incorporates ideas of something being a substance or quality of that substance, ideas of one thing causing another, or one thing being related by necessity or by accident to another. In short, the raw data of sensory input is only a small part of what constitutes human knowledge. Most of it is contributed by the human mind itself; and, so far as human knowledge is concerned, rather than the mind trying to accommodate itself to the external world, the world conforms to the requirements of human sensibility and rationality. Kant compared his radical reorientation of the way philosophers ought to study human knowledge to the Copernican revolution in astronomy. Just as the Earth revolves around the Sun, contrary to common sense, objects conform themselves to the human mind, contrary to common sense.

Kant's idealism notwithstanding, he also believed that a world existed independent of the human mind and completely unknowable by it. This world consists of things-in-themselves, which do not exist in space and time, are not organized in causal relations, and so on, because these are elements contributed by the human mind as conditions for knowing. Because of his commitment to realism (minimal though it may be) Kant was disturbed by Berkeley's uncompromising idealism, which amounted to a denial of the external world. Kant found this incredible and rejected "the absurd conclusion that there can be appearance without anything that appears."

Kant's goal, as developed in *Critique of Pure Reason* (1781), was to supplant Berkeley's crude idealism with a transcendental idealism. The difference, as Kant saw it, is that, while Berkeley began empirically by noting that everything that humans are rationally justified in asserting to exist is related to consciousness, he went on to ask what necessary conditions underlie any empirical experience at all. Kant did not deny that there is empirical experience, but he was critical of Berkeley for not excavating its rational underpinnings. Kant is called a "rationalist" because he thought that the conditions for empirical experience can only be reasoned to, not discovered in, experience; he called his idealism "transcendental" because the conditions he was looking for are common to—they transcend—any experience. In his notorious "proof of an external world," he claimed that he experienced himself as an object in time, that time requires something permanent outside of his consciousness as a precondition for his existence in time, and hence that an external world exists. In other words, the claim is that inner experience presupposes an outer or external world. But few philosophers have claimed to understand why this should be so, and the very contrast of inner and outer seems to beg the question.

Transcendental idealism

Kant believed that all objects of sensation must be experienced within the limits of space or time. Thus, all physical objects have a spatiotemporal location. Because space and time are the backdrop for all sensations, he called them pure forms of sensibility. In addition to these forms, there are also pure forms of understanding, that is, categories or general structures of thought that the human mind contributes in order to understand physical phenomena. Thus, every empirical object is thought to have some cause, to be either a substance or part of some substance, and so on. The structure of judgments finally leads to the question of what properties the propositions that express judgments (or knowledge) have.

From a logical point of view, the propositions that express human knowledge can be divided according to two distinctions. First is the distinction between propositions that are a priori, in the sense that they are knowable prior

to experience, and those that are a posteriori, in the sense that they are knowable only after experience. Second is the distinction between propositions that are analytic, that is, those in which the predicate is included in the subject, and those that are synthetic, that is, those in which the predicate is not included in the subject. Putting the terms of these two distinctions together yields a fourfold classification of propositions. (1) Analytic a priori propositions include "All bachelors are unmarried" and "All squares have four sides." (2) Analytic a posteriori propositions do not exist, according to Kant, because, if the predicate is conceptually included in the subject, the appeal to experience is irrelevant and unnecessary. Also, the negation of an analytic proposition is a contradiction; but, because any experience is contingent, its opposite is logically possible and hence not contradictory. (3) Synthetic a priori propositions include "Every event has a cause" and "7 + 5 = 12." Although it is not part of the concept of an event that it be a cause, it is universally true and necessary that every event has a cause. And, because 12 is a different concept from seven, five, and plus, it does not include any of them singly or jointly as a part of it. (4) Finally, synthetic a posteriori propositions include, "The cat is on the mat" and "It is raining." They are straightforwardly and uncontroversially empirical propositions that are not necessary and are discoverable through observation.

Kant's view that human experience is bounded by space and time and that it is intelligible only as a system of completely determined causal relations existing between events in the world and not between the world and anything outside of it has the consequence that there can be no knowledge of God, freedom, or human immortality. Each of these ideas exceeds the bounds of empirical experience and hence is banished from the realm of reason. As he said, he "found it necessary to deny *knowledge,* in order to make room for faith."

G.F.W. Hegel. G.F.W. Hegel (1770–1831) developed his epistemology pari passu with ontology. Since his positive views are difficult and replete with technical terms, his epistemology is not susceptible of summary here. Some of his criticisms of earlier epistemological views, however, should be mentioned since they helped to bring modern philosophy to a close.

Empiricism takes cognition of particular sensed objects as the foundation for knowledge. But, Hegel argues, no sensation is purely particular. For every sensation consists of something that has a certain feature, quality, or feel, and this feature, quality, or feel is something common to other sensations and hence not particular. Also, all knowledge must be expressible in language, and all fully articulated language uses predicates, which express concepts. Even if the empiricist attempts to represent his knowledge with a single, purely demonstrative word, say, "this" or "now," his view is contradictory. For "this" is common to any indicated object, and "now" can be used to refer to any time. An analogous argument holds against anyone who, like Descartes or Kant, wants to begin with the referent of "I."

Another mistake common to empiricism and rationalism is to think that knowledge requires a correspondence between a person's beliefs and reality. The search for such correspondence is logically absurd since every such search ends with some belief about whether the correspondence holds or not, and thus one has not advanced beyond belief. Kant's distinction between the thing-in-itself and the phenomenon of consciousness is an instance of this absurdity. To make the distinction is to have the object in itself in consciousness and hence not in itself. Thus, Hegel concludes that knowledge and reality cannot be two things but must be identical. Knowledge cannot be perspectival or relative to each person; it is as absolute and objective as reality.

CONTEMPORARY PHILOSOPHY

Contemporary philosophy begins in the late 19th and early 20th century. Much of what sets contemporary philosophy off from modern philosophy is its explicit criticism of the modern tradition and sometimes its apparent indifference to it. There are two basic strains of contemporary philosophy: Continental philosophy, which designates the philosophical style of western European philosophers, and Anglo-American, or analytic, philosophy, which includes the work of many European philosophers who immigrated to Britain, the United States, and Australia shortly before World War II.

Continental philosophy. In epistemology, Continental philosophers during the first quarter of the 20th century were preoccupied with the problem of overcoming the apparent gap between the knower and the known. If a human being has access only to his own ideas of the world and not the world itself, how can there be knowledge at all?

The German philosopher Edmund Husserl (1859–1938) thought that the standard epistemological theories had become intrusive because philosophers were attending to repairing or complicating them rather than focusing on the phenomena of knowledge as humans experience them. To emphasize this reorientation of thinking, he adopted the slogan, "To the things themselves." Philosophers needed to recover the sense of what is given in experience itself, and this could only be accomplished through a careful description of phenomena. Thus, Husserl called his philosophy "phenomenology," which was to begin as a purely descriptive science and only later to ascend to a theoretical, or "transcendental," science. Phenomenology

Husserl thought that the philosophies of Descartes and Kant presupposed a gap between the aspiring knower and what is known and that the experience of the external world was thus dubious and had to be proven. These presuppositions violated Husserl's belief that philosophy, as the most fundamental science, should be free of presuppositions. Thus, he held that it is illegitimate to assume there to be any problem of knowledge or of the external world prior to an investigation of the matter without any presuppositions. Husserl's device to cut through the Gordian knot of such assumptions was to introduce an "*epochē*." In other words, he would bracket or refuse to consider traditional philosophical problems until after the phenomenological description had been completed.

The *epochē* was just one of a series of so-called transcendental reductions that Husserl proposed in order to ensure that he was not presupposing anything. One of these reductions supposedly gave one access to "the transcendental ego," or "pure consciousness." Although one might expect phenomenology then to describe the experience or contents of this ego, Husserl instead aimed at "eidetic reduction," that is, the discovery of the essences of various sorts of ideas, such as redness, surface, or relation. All of these moves were part of Husserl's desire to discover the one, perfect methodology for philosophy in order to ensure absolute certainty.

Because Husserl's transcendental ego seems very much like the Cartesian mind that thinks of a world but does not have either direct access to or certainty of it, Husserl tried in *Cartesianische Meditationen* (1931; "Cartesian Meditations") to overcome the apparent gap, the very thing he had set out either to destroy or bypass. Because the transcendental ego seems to be the only genuinely existent consciousness, Husserl also tried to overcome the problem of solipsism.

Many of Husserl's followers, including his most famous student, Martin Heidegger (1889–1976), recognized that something had gone radically wrong with the original direction of phenomenology. According to Heidegger's diagnosis, the root of the problem was Husserl's assumption that there is an "Archimedean point" for human knowledge, to use Husserl's own phrase; but, there is no ego detached from the world and filled with ideas or representations, according to Heidegger. In *Being and Time* (1927) Heidegger returned to the original formulation of the phenomenological project as a return to the things themselves. Thus, all the transcendental reductions are abandoned. What he claimed to discover is that human beings are inherently world-bound. The world does not need to be derived; it is presupposed by human experience. In their prereflective experience, humans inhabit a sociocultural environment, in which the primordial kind of cognition is practical and communal, not theoretical or individual ("egoistic"). Human beings interact with the

things of their everyday world (*Lebenswelt*) as a workman interacts with his tools; they hardly ever approach the world as a philosopher or scientist would. The theoretical knowledge of a philosopher is a derivative and specialized form of cognition, and the major mistake of epistemology from Descartes to Kant to Husserl was to take philosophical knowledge as the paradigm for all knowledge.

Heidegger's insistence that a human being is something that inhabits a world notwithstanding, he marked out human reality as ontologically special. He called this reality *Dasein,* the being, apart from all others, which is present to the world. Thus, like the transcendental ego, a cognitive being takes pride of place in Heidegger's philosophy.

In France the principal phenomenological proponent of the mid-century was Maurice Merleau-Ponty (1908–61). But he rejected Husserl's bracketing of the world, that is, his mistake in not recognizing that human experience of the world is primary, a view capsulized in Merleau-Ponty's phrase "the primacy of perception." He furthermore held that dualistic analyses of knowledge, such as the Cartesian mind–body dualism, are inadequate. In fact, no conceptualization of the world can be complete in his view. Because human cognitive experience requires a body and the body a position in space, human experience is necessarily perspectival and thus incomplete. Although humans experience material beings as multidimensional objects, part of the object always exceeds the cognitive grasp of the person just because of his limited perspective. In *Phenomenology of Perception* (1945), Merleau-Ponty develops these ideas (along with a detailed attack on the sense-datum theory, discussed below).

The epistemological views of Jean-Paul Sartre (1905–80) share some features with Merleau-Ponty's. Both reject Husserl's transcendental reductions, and both think of human reality as being-in-the-world. But Sartre's views have Cartesian elements that were anathema to Merleau-Ponty. Sartre distinguished between two basic kinds of being. Being-in-itself (*en soi*) is the inert and determinate world of nonhuman existence. Over and against it is being-for-itself (*pour soi*), which is the pure consciousness that defines human reality.

Later Continental philosophers attacked the entire philosophical tradition from Descartes to the 20th century for its explicit or implicit dualisms. Being/nonbeing, mind/body, knower/known, ego/world, being-in-itself/being-for-itself are all variations on a way of philosophizing that the philosophers of the last third of the 20th century have tried to undermine. The structuralist Michel Foucault (1926–84) wrote extensive historical studies, most notably *The Archaeology of Knowledge* (1969), in order to demonstrate that all concepts are historically conditioned and that many of the most important ones serve the political function of controlling people rather than any purely cognitive purpose. Jacques Derrida has claimed that all dualisms are value-laden but indefensible. His technique of "deconstruction" attempts to show that every philosophical dichotomy is incoherent, because whatever can be said about one term of the dichotomy can also be said of the other.

Michel Foucault

Dissatisfaction with the Cartesian philosophical tradition can also be found in the United States. The American pragmatist John Dewey (1859–1952) directly challenged the idea that knowledge is primarily theoretical; experience, he argued, consists of an interaction between a living being and his environment. Knowledge is not a fixed staring at something but a process of acting and being acted upon. Richard Rorty has done much to reconcile Continental and Anglo-American philosophy. He has argued that Dewey, Heidegger, and Ludwig Wittgenstein are the three greatest philosophers of the 20th century, specifically because of their attacks on the epistemological tradition of modern philosophy. (A.P.Ma.)

Analytic philosophy. Analytic philosophy, the prevailing philosophy in the Anglo-American world in the 20th century, has its origins in symbolic logic on the one hand and in British empiricism on the other. Some of its important contributions have been nonepistemological in character, but in the area of epistemology its contributions have also been of the first order. Its main characteristics have been the avoidance of system building and a commitment to detailed, piecemeal analyses of specific issues. Within this tradition there have been two main approaches: a formal style, deriving from logic; and an approach emphasizing ordinary language. Among those who can be identified with the first method are Bertrand Russell, Gottlob Frege, Rudolf Carnap, Alfred Tarski, and W.V.O. Quine; and among those with the second are G.E. Moore, Gilbert Ryle, J.L. Austin, Norman Malcolm, P.F. Strawson, and Zeno Vendler. Wittgenstein can be situated in both groups, his early work belonging to the former tradition and his posthumous works, *Philosophical Investigations* (1953) and *On Certainty* (1969), to the latter.

Perhaps the most distinctive feature of analytic philosophy is its emphasis upon the role that language plays in the creation and resolution of philosophical problems. These problems, it is said, arise through the misuses, oversimplifications, and unwarranted generalizations of everyday language. Wittgenstein said in this connection: "Philosophy is a battle against the bewitchment of the intelligence by means of language." The idea that philosophical problems are in some important sense linguistic (or conceptual) is called the "linguistic turn."

The linguistic turn

Commonsense philosophy, logical positivism, and naturalized epistemology. Three of the most notable achievements of analytic philosophy are commonsense philosophy, logical positivism, and naturalized epistemology. G.E. Moore (1873–1958) made a defense of what he called the commonsense view of the world. According to Moore, virtually everybody knows certain propositions to be true, such as that the Earth exists, that it is very old, and that other persons now exist on it. Furthermore, any philosophical theory that runs counter to this commonsense view can be rejected out of hand as mistaken. All forms of idealism fall into this category. Wittgenstein, for whom certainty is that "which stands fast for all of us," extended this view. In *On Certainty* he argued that certitude is connected with action and that "Action lies at the bottom of the language game."

The development of logical positivism (also called logical empiricism) was a product of the Vienna Circle under the leadership of the German logical empiricist philosopher Moritz Schlick, and it became a dominant form of philosophy in England with the publication of A.J. Ayer's *Language, Truth, and Logic* (1936). Logical positivism holds that all significant propositions are either those of logic or mathematics on the one hand or those of science on the other. Since the utterances of traditional philosophy (especially metaphysics) fall into neither of these groups, they are unverifiable in principle and accordingly can be rejected as nonsense. The only legitimate function for philosophy is conceptual analysis, *i.e.,* the clarification of various notions, such as "probability" or "causality."

W.V.O. Quine (b. 1908), in "Two Dogmas of Empiricism" (1950), launched an attack upon the notion that there is a difference in kind between analytic and synthetic statements. Quine argued powerfully that the so-called difference is one of degree. In a later work, *Word and Object* (1960), Quine developed a new type of philosophy, which he called "naturalized epistemology." He rejected the notion that epistemology has a normative function and claimed that its only legitimate role is to describe the way knowledge is actually obtained. In effect, its function is to describe how present science arrives at the beliefs accepted by the scientific community.

Perception and knowledge. To a great extent the epistemological interests of analytic philosophers in the 20th century have been concentrated upon the relationship between knowledge and perception. The major figures in this development have been Bertrand Russell, G.E. Moore, H.H. Price, C.D. Broad, A.J. Ayer, and H.P. Grice. Although their views differed considerably—Russell, Broad, and Ayer were phenomenalists, Grice was a defender of the causal theory of perception, and Moore attempted to construct a theory of direct realism—all of them were defenders of sense-data theory (see below).

Sense-data theory was criticized by proponents of the so-called theory of appearing, such as G.A. Paul and W.H.F. Barnes, who claimed that the arguments for the existence

of sense-data are spurious. Those arguments assume, for example, that because a penny looks elliptical from a certain perspective, it follows that there exists an elliptical object (sense-datum), which an observer is directly apprehending. They denied the inference, saying that the introduction of a separate entity, a sense-datum, does not follow from the fact that a circular object looks elliptical and to believe that it does is simply to misdescribe certain common perceptual situations. The most powerful attack on sense-data theory was generated by J.L. Austin in *Sense and Sensibilia* (1962).

Many philosophers, in turn, rejected the theory of appearing. They felt that puzzles about the status of illusions and other visual anomalies still require explanation. Their aim was to give a coherent account of how knowledge is possible despite the existence of perceptual error. Realism and phenomenalism are the two main types of theories developed to account for these difficulties.

Both realism and phenomenalism have had numerous variants. Two forms of realism, direct (naive) realism and representative realism (also called "the causal theory"), are historically important.

Realism. Realism is both a metaphysical and an epistemological theory. The realist is committed to two principles: first, that some of the objects apprehended through perception are public and, second, that some of those objects are mind-independent. It is especially the second of these notions that distinguishes realists from phenomenalists.

The realist believes that there is an intuitive commonsense distinction among various classes of entities perceived by human beings. One class consists, among others, of headaches, thoughts, pains, or desires, and the other of tables, rocks, planets, persons, animals, and certain physical phenomena such as rainbows, lightning, and shadows. The metaphysical aspect of realism sees the former as mental, the latter as physical. A realist metaphysics maintains that the classes are mutually exclusive. What a realist epistemology adds to this metaphysics is that mental entities are private, whereas physical objects are public. By "private" it is meant that each item belonging to the category of the mental is apprehensible by one person only. Thus, only one person can have a particular headache or a particular pain. In contrast, physical objects are public; more than one person can see or touch the same chair.

The realist also believes that items belonging to the class of the physical are mind-independent. What is meant by this notion is that the existence of these objects does not depend upon their being perceived by anyone. Thus, whether or not a particular table is being seen or touched by someone has no effect upon its existence. Even if nobody is looking at it, it would still exist (other things being equal). But this is not true of mental phenomena. If somebody is not actually having a headache, realists would deny that the headache exists. A headache is thus mind-dependent in a way in which tables, rocks, and shadows are not.

Realist theories of knowledge thus begin by assuming the public–private distinction, and most realists start by assuming that one does not have to prove the existence of mental phenomena. These are things of which each person is directly aware, and there is no special "problem" about their existence. But they do not assume this to be true of physical phenomena. As the existence of visual aberrations, illusions, and other anomalies shows, one cannot be sure that in any perceptual situation one is apprehending physical objects. All a person can be sure of is that he is aware of something, an appearance of some sort, say of a bent stick in water; but whether that appearance corresponds to anything actually existing in the external world is an open question.

In the *Foundations of Empirical Knowledge* (1940) Ayer called this difficulty "the egocentric predicament." When a person looks at what he thinks is a physical object, such as a chair, what he is directly apprehending is a certain visual appearance. But such an appearance seems to be private to that person; it seems to be something mental and not publicly accessible. What then justifies the individual's belief in the existence of supposedly external objects—*i.e.*, physical entities that exist external to the human mind? Direct realism and representative realism are the two main theoretical responses to this challenge.

Both direct realism and representative realism rely strongly on sense-data theory. The technical term "sense-datum," which played an important role in the development of versions of both theories, is sometimes explained by using examples. If one is hallucinating and sees pink rats, one is seeing a sense-datum. Although no real rats are there, one is having a certain visual sensation as of coloured rats, and this sensation is what is called a sense-datum. The image one sees with one's eyes closed after looking fixedly at a bright light is another example. But, even in normal vision, one can be said to be apprehending sense-data. For instance, in looking at a round penny from a certain angle, one will see the penny to be elliptical. In such a case, there is an elliptical sense-datum in one's visual field. This last example was held by Broad, Price, and Moore to be particularly important, for it makes a strong case for holding that one always sees sense-data, whether perception is normal or abnormal.

Sense-data theory

According to defenders of sense-data theory, what these examples have in common is that in every perceptual act one is directly aware of something. A sense-datum is thus frequently defined as an entity that is the object of direct perception. By "direct" these philosophers mean that no inference is necessary in order to apprehend these entities. According to Broad, Price, and Ayer, sense-data differ from physical objects in having the properties they appear to have; *i.e.*, they cannot appear to have properties they do not really have. The problem for a realist who accepts sense-data is to show how these private sensations allow justification of the intuitive belief that there are physical objects which exist outside of the individual's perception. Russell in particular tried to show in such works as *The Problems of Philosophy* (1912) and *Our Knowledge of the External World* (1914) how knowledge of the external world could be built up from such mental, private apprehensions.

Direct realism

During the 20th century direct realism took many forms; indeed there were direct realists, such as James J. Gibson who, in *The Ecological Approach to Visual Perception* (1979), rejected sense-data theory and claimed that the outside aspects (the physical surfaces) of physical objects are normally directly observed. But many realists, such as G.E. Moore and his followers, believed that the existence of sense-data must be accepted. Moore took the unusual step of suggesting that such sense-data might not be mental entities but could be a physical part of the surface of the perceived material object. Thompson Clarke in "Perceiving Physical Objects and Surfaces" (1965) went beyond Moore in arguing that one normally directly perceives the whole physical object itself.

All of these views have problems in dealing with perceptual anomalies. In fact, Moore, in his last published paper, "Visual Sense-Data" (1957), abandoned the attempt to defend direct realism. He held that, because the elliptical sense-datum one perceives when one looks at a round coin cannot be identical with the circular surface of the coin, one cannot be seeing the coin directly but only the sense-datum. Hence, one cannot have direct knowledge of external objects.

Because of the problems associated with direct realism, many philosophers, including H.H. Price, H.P. Grice, and Robert E. French, have argued for the causal theory, that is, the theory of representative realism. This is an old view whose most famous exponent in early modern philosophy was Locke. It is also sometimes called "the scientific theory" because it seems to be supported by findings in optics and physics. According to this form of realism there are real physical objects that exist external to the human mind, and there are also sense-data (or their equivalents, such as so-called mental representations). Visual perception is then explained as follows. Light is reflected from external objects, moves through space according to well-known laws of physics, is picked up by the human visual system, which includes the eye, the optic nerve, and the retina, and then is ultimately processed by the brain. This is a causal sequence. Light causes a reaction in the eye, that

Representative realism

reaction is the cause of a response in the optic nerve, and so forth. The last event in this causal sequence is "seeing."

What one is apprehending in such a case is a mental representation (sense-datum) of the original object; and, through various processes in the brain, this representation gives human beings a depiction of the object as it is. Visual illusion is explained in various ways, but usually as the result of some anomaly in the causal chain that gives rise to distortions and other types of aberrant visual phenomena. In such a view, human observers are directly aware of mental representations, or sense-data, and only indirectly aware of the physical objects that cause these data in the brain.

The difficulty with this view is that, since one cannot compare the sense-datum that is directly perceived with the original object, one cannot ever be sure that it gives an accurate representation of it; and therefore human beings cannot know that the real world corresponds to their perception of it. They are still confined within the circle of appearance after all. It thus seems that neither version of realism satisfactorily solves the problem it began with.

Phenomenalism. In light of these difficulties with realist theories of perception some philosophers, so-called phenomenalists, proposed a completely different way of analyzing the relationship between perception and knowledge. In particular, they rejected the distinction between independently existing physical objects and mind-dependent sense-data that direct realism presupposes. They claimed that either the very notion of an independent existence is nonsense because human beings have no evidence for it or that what is meant by "independent existence" must be reinterpreted in such a way as not to go beyond the sort of perceptual evidence human beings do or could have for the existence of things. In effect, these philosophers challenged the cogency of the intuitive ideas that the ordinary person supposedly has about independent existence.

All variants of phenomenalism are strongly verificationist in thrust. That is, they wish to maintain that belief in an external world must be capable of verification or confirmation, and this entails that such a belief cannot be acceptable if it goes beyond the realm of possible perceptual experience.

Phenomenalists have thus tried to analyze in wholly perceptual terms what it means to say that any object, say a tomato, exists. They claim that any such analysis must start by deciding what is meant by a tomato. In their view a tomato is something that has certain properties, including a certain size, weight, colour, and shape. If one were to abstract the total set of such observed properties from the object, nothing would be left over; there would be no presumed Lockean "substratum" that supports these properties and which is itself unperceived. There is thus no evidence in favour of such an unperceivable feature, and no reference to it is needed in explaining what a tomato or any so-called physical object is.

To talk about any existent object is thus to talk about a collection of perceivable features localized in a particular portion of space-time. Hence, what one means by a tomato is something that in principle must be perceivable. Accordingly, to say that a tomato exists is either to describe a collection of properties that an observer is actually perceiving or a collection that such an observer would perceive under certain specified conditions. To say, for instance, that a tomato exists in the next room is to say that, if one went to that room, one would see a familiar reddish shape, would obtain a certain taste if one bit into it, or would feel something soft and smooth if one touched it. To speak about that tomato's existing unperceived in the next room thus does not entail that it is unperceivable. In principle, everything that exists is perceivable. Therefore, the notion of existing independently of perception has been misunderstood or mischaracterized by both philosophers and nonphilosophers. Once it is understood that objects are merely sets of properties and that such collections of properties are in principle always perceivable, the notion that there is some sort of unbridgeable gap between people's perceptual evidence and the existence of an object is just a mistake, a confusion between the concepts of actually being perceived and of being perceivable.

In this view, perceptual error is explained in terms of coherence and predictability. To say with truth that one is perceiving a tomato means that one's present set of perceptual experiences and an unspecified set of future experiences will "cohere." That is, if the object a person is looking at is a tomato, then he can expect that, if he touches, tastes, and smells it, he will receive a recognizable grouping of sensations. If the object he has in his visual field is hallucinatory, then there will be a lack of coherence between what he touches, tastes, and smells. He might see a red shape but not be able to touch or taste it.

The criterion of coherence

The theory is generalized to include what others would touch, see, and hear as well, so that what the realists call "public" will also be defined in terms of the coherence of perceptions. A so-called physical object is public if the perceptions of many persons cohere or agree, and otherwise it is not. This explains why a headache is not a public object. In similar fashion, a so-called physical object will be said to have an independent existence if the expectations of future perceptual experiences are borne out. If tomorrow, or the day after, a person has similar perceptual experiences to those he had today, then he can say that the object he is perceiving has an independent existence. The phenomenalist thus attempts to account for all the facts that the realist wishes to explain without positing the existence of anything that transcends possible experience.

The criticisms of this view tend to be technical. Generally speaking, however, realists have objected to it on the ground that it is counterintuitive to think of a tomato as being a set of actual or possible perceptual experiences. The realist argues that human beings do have such experiences, or under certain circumstances would have them, because there is an object out there that exists independently of them and is their source. Phenomenalism, they contend, has the implication that, if no perceivers existed, then the world would contain no objects; and, if this is a consequence of the view, then it is surely inconsistent both with what ordinary persons believe and with the known scientific fact that all sorts of objects existed in the universe long before there were any perceivers. But its supporters deny that phenomenalism carries such an implication, and the debate about its merits remains unresolved.

PHILOSOPHY OF MIND AND EPISTEMOLOGY

In the late 1970s a series of developments occurred in a variety of intellectual fields that promise to cast new light on the nature of the human mind. There have been explosive advances in neuroscience, psychology, cognitive science, neurobiology, artificial intelligence, and computer studies. These have resulted in a new understanding of how seeing works, how the mind forms representations of the external world, how information is stored and retrieved, and the ways in which calculations, decision procedures, and other intellectual processes resemble and differ from the operations of sophisticated computers, especially those capable of parallel processing.

The implications for epistemology of these developments are equally exciting. They promise to give philosophers new understandings of the relationship between common sense and theorizing, that is, whether some form of materialism which eliminates reference to mental phenomena is true or whether the mental–physical dualism which common sense assumes is irreducible, and they also open new avenues for dealing with the classical problem of other minds. It is too early to make an assessment of the relevance for epistemology of what has already been achieved in these areas. There is no doubt, however, that these advances are revolutionary and that a new area of intellectual discovery has begun. (Av.S.)

BIBLIOGRAPHY. The texts of the classics mentioned below for which specific editions have not been noted are available in many English-language translations; two notable collections are *The Loeb Classical Library* and *Oxford Classical Text* series.

The history of epistemology: (Ancient): An excellent collection on skepticism is MILES BURNYEAT (ed.), *The Skeptical Tradition* (1983). For Greek Skepticism in particular, see CHARLOTTE L. STOUGH, *Greek Skepticism: A Study in Epistemology* (1969). The chief epistemological works of PLATO are his *Meno, Theaetetus,* and *Republic,* especially Books V–VII. The views

of ARISTOTLE can be found in *On the Soul, Metaphysics,* Book IV, ch. 5 and 6, and *Posterior Analytics,* Book I, ch. 3. The locus classicus for ancient skepticism is R.G. BURY (trans.), *Sextus Empiricus,* 4 vol. (1933–49), in *The Loeb Classical Library* series. From among the voluminous writings of AUGUSTINE, see *Against the Academicians,* trans. by MARY PATRICIA GARVEY (1942, reissued 1978).

(*Medieval*): For the period as a whole, see appropriate articles in *The Cambridge History of Later Greek and Early Medieval Philosophy,* ed. by A.H. ARMSTRONG (1967); and *The Cambridge History of Later Medieval Philosophy: From the Rediscovery of Aristotle to the Disintegration of Scholasticism, 1100–1600,* ed. by NORMAN KRETZMANN, ANTHONY KENNY, and JAN PINBORG (1982). For the thoughts of ANSELM OF CANTERBURY, see his *Proslogium,* ch. 1, and *On Truth.* THOMAS AQUINAS, *Summa Theologiae,* discusses the soul in general in part I, question 77, and the intellectual powers of the soul in part I, question 79. The views of JOHN DUNS SCOTUS can be found in the relevant sections of his *Philosophical Writings,* trans. by ALLAN WOLTER (1962, reprinted 1987); and in A.P. MARTINICH, "Duns Scotus on the Possibility of an Infinite Being," *Philosophical Topics,* supplementary vol. 80, pp. 23–29 (1982). The ideas of WILLIAM OF OCKHAM can be found in the relevant sections of his *Philosophical Writings,* trans. by PHILOTHEUS BOEHNER (1957, reissued 1967).

(*Modern*): Two excellent and now classic histories of early modern philosophy from different perspectives are EDWIN ARTHUR BURTT, *The Metaphysical Foundations of Modern Physical Science,* rev. ed. (1972), which emphasizes the effect of modern science on philosophy; and RICHARD H. POPKIN, *The History of Scepticism from Erasmus to Spinoza,* rev. and expanded ed. (1979), which emphasizes the rediscovery of skepticism in the 16th century. RENÉ DESCARTES's greatest work is *Meditations on First Philosophy,* trans. by JOHN COTTINGHAM (1986; originally published in Latin, 1641). JOHN LOCKE attacks the doctrine of innate ideas in Book I of his *An Essay Concerning Human Understanding,* ed. by PETER H. NIDDITCH (1975), while his position on knowledge is developed in Books II and IV. A good introduction to Locke's thought is JOHN W. YOLTON, *Locke: An Introduction* (1985). The best work of GEORGE BERKELEY is *A Treatise Concerning the Principles of Human Knowledge,* ed. by KENNETH WINKLER (1982); a more popular presentation of his views is *Three Dialogues Between Hylas and Philonous,* ed. by ROBERT MERRIHEW ADAMS (1979). DANIEL E. FLAGE, *Berkeley's Doctrine of Notions: A Reconstruction Based on His Theory of Meaning* (1987), discusses a central but neglected aspect of Berkeley's epistemology. DAVID HUME's most expansive discussion of knowledge is in Book I of *A Treatise of Human Nature,* 2nd ed., edited by L.A. SELBY-BIGGE and rev. by P.H. NIDDITCH (1978). A later and more accessible statement of Hume's view is presented in *An Enquiry Concerning Human Understanding,* ed. by ERIC STEINBERG (1977). IMMANUEL KANT, *Critique of Pure Reason,* trans. by NORMAN KEMP SMITH (1929, reissued 1978; originally published in German, 1781), is Kant's greatest work. The best clear, brief, and accurate explanation of Kant's epistemology is A.C. EWING, *A Short Commentary on Kant's "Critique of Pure Reason"* (1938, reprinted 1987). An important book that rejects the view of Kant as a phenomenalist or subjective idealist is HENRY E. ALLISON, *Kant's Transcendental Idealism: An Interpretation and Defense* (1983). For G.F.W. HEGEL's criticisms of Kant, see his *Lectures on the History of Philosophy,* trans. from German by ELIZABETH S. HALDANE and FRANCES H. SIMSON (1896, reprinted 1974), part III, section iii, B. A major study on the relationship between Kant and Hegel is ROBERT B. PIPPIN, *Hegel's Idealism: The Satisfactions of Self-Consciousness* (1989).

(*Contemporary*): A short and readable history of Continental philosophy is ROBERT C. SOLOMON, *Continental Philosophy Since 1750: The Rise and Fall of the Self* (1988). MARTIN HEIDEGGER, *Being and Time,* trans. by JOHN MACQUARRIE and EDWARD ROBINSON (1962, reissued 1978; originally published in German, 1927), presents an alternative epistemological scheme. JOHN DEWEY, *The Quest for Certainty: A Study of the Relation of Knowledge and Action* (1929, reissued 1979), is an attack on modern epistemology by an American pragmatist. RICHARD RORTY, *Philosophy and the Mirror of Nature* (1979), a history of modern and contemporary philosophy, has attracted great attention as an attack on classical epistemology from an analytically trained philosopher. PAUL FEYERABEND, *Against Method,* 2nd ed. (1988), advocates what he describes as "an anarchistic theory of knowledge." KARL R. POPPER, *Objective Knowledge: An Evolutionary Approach,* rev. ed. (1979), argues against a tradition that goes back at least to Aristotle and rejects the subjective interpretation of knowledge, according to which knowledge is located in individual people.

The 20th-century literature on perception and knowledge is vast. A good general collection is ROBERT J. SWARTZ (ed.), *Perceiving, Sensing, and Knowing* (1965, reissued 1976), which includes two important attacks upon sense-data theory—W.H.F. BARNES, "The Myth of Sense-Data," and G.A. PAUL, "Is There a Problem About Sense-Data?"—and a strong defense of the sense-datum view by C.D. BROAD, "The Theory of Sensa." The most important pre-World War II books on perception and knowledge are BERTRAND RUSSELL, *The Problems of Philosophy* (1911, reissued 1988), and *Our Knowledge of the External World* (1926, reissued 1972); G.E. MOORE, *Philosophical Studies* (1922, reissued 1970), especially the important articles defending sense-data theory, "Some Judgments of Perception" and "The Status of Sense-Data"; H.H. PRICE, *Perception* (1932, reprinted 1981), which invokes the notion of a sense-datum in defense of the causal theory of perception; and ALFRED J. AYER, *The Foundations of Empirical Knowledge* (1940, reissued 1971), which merges sense-data theory with the principles of logical positivism. Notable works since World War II include GILBERT RYLE, *The Concept of Mind* (1949, reprinted 1984), which defends a sophisticated form of epistemological behaviourism; RODERICK M. CHISHOLM, *Perceiving: A Philosophical Study* (1957); and J.L. AUSTIN, *Sense and Sensibilia* (1962), which contains a withering assault on the sense-data theory from the standpoint of ordinary-language philosophy. Surfaces, perception, and knowledge are discussed in THOMPSON CLARKE, "Seeing Surfaces and Physical Objects," in MAX BLACK (ed.), *Philosophy in America* (1965); JAMES J. GIBSON, *The Ecological Approach to Visual Perception* (1979, reissued 1986); and AVRUM STROLL, *Surfaces* (1988). The theory of representative realism is given a sophisticated defense in FRANK JACKSON, *Perception* (1977); and S. ULLMAN, "Against Direct Perception," *Behavioral & Brain Sciences,* 3(3):373–415 (September 1980), attacks direct realism, especially J.J. Gibson's version of that theory, from a standpoint of modern cognitive science. A comprehensive survey of the literature from about 1980 to 1984 on direct realism and representative realism is to be found in EDMOND WRIGHT, "Recent Work in Perception," *American Philosophical Quarterly,* 21:17–30 (January 1984).

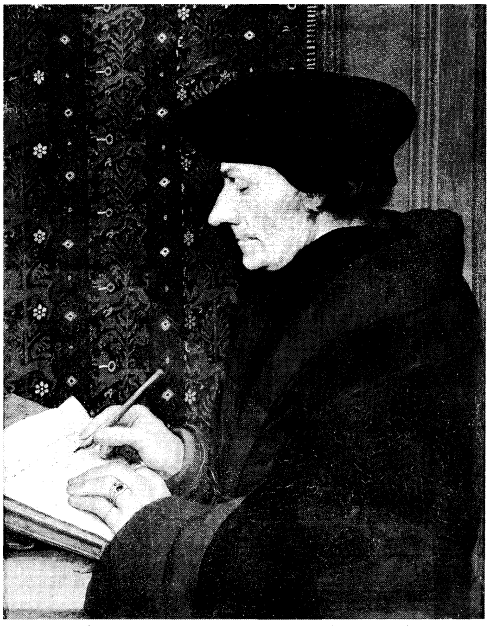
Knowledge and the commonsense view of the world are discussed by G.E. MOORE, "A Defence of Common Sense," in his *Philosophical Papers* (1959); LUDWIG WITTGENSTEIN, *On Certainty,* trans. from German (1969); NORMAN MALCOLM, *Thought and Knowledge* (1977); and JOHN R. SEARLE, *Intentionality: An Essay in the Philosophy of Mind* (1983). An excellent simple survey of the impact of computer studies, work in artificial intelligence, neuroscience, and neurobiology on our knowledge of other minds is found in PAUL M. CHURCHLAND, *Matter and Consciousness: A Contemporary Introduction,* rev. ed. (1988).

Two excellent anthologies are HAROLD MORICK (ed.), *Challenges to Empiricism* (1972, reprinted 1980); and PAUL K. MOSER and ARNOLD VANDER NAT (eds.), *Human Knowledge: Classical and Contemporary Approaches* (1987). EDMUND L. GETTIER, "Is Justified True Belief Knowledge?" in *Analysis,* 23:121–23 (June 1963), is considered by many to be a decisive refutation of the justified, true-belief analysis of knowledge. NOAM CHOMSKY, *Language and the Problems of Knowledge* (1988), discusses innateness, language, and psychology. RODERICK M. CHISHOLM, *Theory of Knowledge,* 3rd ed. (1989); and ROBERT AUDI, *Belief, Justification, and Knowledge* (1988), are two good introductions to standard epistemological problems. (A.P.Ma./Av.S.)

Erasmus

Born in Rotterdam in 1469, Desiderius Erasmus was the greatest European scholar of the 16th century. Using the philological methods pioneered by Italian humanists, he helped lay the groundwork for the historical-critical study of the past, especially in his studies of the Greek New Testament and the Church Fathers. His educational writings contributed to the replacement of the older scholastic curriculum by the new humanist emphasis on the classics. By criticizing ecclesiastical abuses, while pointing to a better age in the distant past, he encouraged the growing urge for reform, which found expression both in the Protestant Reformation and in the Catholic Counter-Reformation. Finally, his independent stance in an age of fierce confessional controversy—rejecting both Luther's doctrine of predestination and the powers that were claimed for the papacy—made him a target of suspicion for loyal partisans on both sides and a beacon for those who valued liberty more than orthodoxy.

Giraudon—Art Resource/EB Inc.



Erasmus, oil painting by Hans Holbein the Younger, 1523. In the Louvre, Paris.

Early life and career. Erasmus was the second illegitimate son of Roger Gerard, a priest, and Margaret, a physician's daughter. He advanced as far as the third-highest class at the chapter school of St. Lebuin's in Deventer. One of his teachers, Jan Synthen, was a humanist, as was the headmaster, Alexander Hegius. The schoolboy Erasmus was clever enough to write classical Latin verse that impresses a modern reader as cosmopolitan.

School years

After both parents died, the guardians of the two boys sent them to a school in 's Hertogenbosch conducted by the Brethren of the Common Life, a lay religious movement that fostered monastic vocations. Erasmus would remember this school only for a severe discipline intended, he said, to teach humility by breaking a boy's spirit.

Having little other choice, both brothers entered monasteries. Erasmus chose the Augustinian canons regular at Steyn, near Gouda, where he seems to have remained about seven years (1485–92). While at Steyn he paraphrased Lorenzo Valla's *Elegantiae,* which was both a compendium of pure classical usage and a manifesto against the scholastic "barbarians" who had allegedly corrupted it. Erasmus' monastic superiors became "barbarians" for him by discouraging his classical studies. Thus, after his ordination to the priesthood (April 1492), he was happy to escape the monastery by accepting a post as Latin secretary to the influential Henry of Bergen, bishop of Cambrai. His *Antibarbarorum liber,* extant from a revision of 1494–95, is a vigorous restatement of patristic arguments for the utility of the pagan classics, with a polemical thrust against the cloister he had left behind: "All sound learning is secular learning."

Erasmus was not suited to a courtier's life, nor did things improve much when the bishop was induced to send him to the University of Paris to study theology (1495). He disliked the quasi-monastic regimen of the Collège de Montaigu, where he lodged initially, and pictured himself to a friend as sitting "with wrinkled brow and glazed eye" through Scotist lectures. To support his classical studies, he began taking in pupils; from this period (1497–1500) date the earliest versions of those aids to elegant Latin—including the *Colloquia* and the *Adagia*—that before long would be in use in humanist schools throughout Europe.

The wandering scholar. In 1499 a pupil, William Blount, Lord Mountjoy, invited Erasmus to England. There he met Thomas More, who became a friend for life. John Colet quickened Erasmus' ambition to be a "primitive theologian," one who would expound Scripture not in the argumentative manner of the scholastics but in the manner of Jerome and the other Church Fathers, who lived in an age when men still understood and practiced the classical art of rhetoric. The impassioned Colet besought him to lecture on the Old Testament at Oxford, but the more cautious Erasmus was not ready. He returned to the Continent with a Latin copy of St. Paul's Epistles and the conviction that "ancient theology" required mastery of Greek.

Thomas More and John Colet

On a visit to Artois, Fr. (1501), Erasmus met the fiery preacher Jean Voirier, who, though a Franciscan, told him that "monasticism was a life more of fatuous men than of religious men." Admirers recounted how Voirier's disciples faced death serenely, trusting in God, without the solemn reassurance of the last rites. Voirier lent Erasmus a copy of works by Origen, the early Greek Christian writer who promoted the allegorical, spiritualizing mode of scriptural interpretation, which had roots in Platonic philosophy. By 1502 Erasmus had settled in the university town of Louvain (Brabant) and was reading Origen and St. Paul in Greek. The fruit of his labours was *Enchiridion militis Christiani* (1503/04; *Handbook of a Christian Knight*). In this work Erasmus urged readers to "inject into the vitals" the teachings of Christ by studying and meditating on the Scriptures, using the spiritual interpretation favoured by the "ancients" to make the text pertinent to moral concerns. The *Enchiridion* was a manifesto of lay piety in its assertion that "monasticism is not piety." Erasmus' vocation as a "primitive theologian" was further developed through his discovery at Park Abbey, near Louvain, of a manuscript of Valla's *Adnotationes* on the Greek New Testament, which he published in 1505 with a dedication to Colet.

Travels to Italy

Erasmus sailed for England in 1505, hoping to find support for his studies. Instead he found an opportunity to travel to Italy, the land of promise for northern humanists, as tutor to the sons of the future Henry VIII's physician. The party arrived in the university town of Bologna in time to witness the triumphal entry (1506) of the warrior pope Julius II at the head of a conquering army, a scene that figures later in Erasmus' anonymously published satiric dialogue, *Julius exclusus e coelis* (written 1513–14). In Venice Erasmus was welcomed at the celebrated printing house of Aldus Manutius, where Byzantine émigrés enriched the intellectual life of a numerous scholarly company. For the Aldine press Erasmus expanded his *Adagia,* or annotated collection of Greek and Latin adages, into a

monument of erudition with over 3,000 entries; this was the book that first made him famous. The adage "Dutch ear" (*auris Batava*) is one of many hints that he was not an uncritical admirer of sophisticated Italy, with its theatrical sermons and its scholars who doubted the immortality of the soul; his aim was to write for honest and unassuming "Dutch ears."

De pueris instituendis, written in Italy though not published until 1529, is the clearest statement of Erasmus' enormous faith in the power of education. With strenuous effort the very stuff of human nature could be molded, so as to draw out (*e-ducare*) peaceful and social dispositions while discouraging unworthy appetites. Erasmus, it would almost be true to say, believed that one is what one reads. Thus the "humane letters" of classical and Christian antiquity would have a beneficent effect on the mind, in contrast to the disputatious temper induced by scholastic logic-chopping or the vengeful amour propre bred into young aristocrats by chivalric literature, "the stupid and tyrannical fables of King Arthur."

The celebrated *Moriae encomium,* or *Praise of Folly,* conceived as Erasmus crossed the Alps on his way back to England and written at Thomas More's house, expresses a very different mood. For the first time the earnest scholar saw his own efforts along with everyone else's as bathed in a universal irony, in which foolish passion carried the day: "Even the wise man must play the fool if he wishes to beget a child."

Little is known of Erasmus' long stay in England (1509–14), except that he lectured at Cambridge and worked on scholarly projects, including the Greek text of the New Testament. His later willingness to speak out as he did may have owed something to the courage of Colet, who risked royal disfavour by preaching a sermon against war at the court just as Henry VIII was looking for a good war in which to win his spurs. Having returned to the Continent, Erasmus made connections with the printing firm of Johann Froben and traveled to Basel to prepare a new edition of the *Adagia* (1515). In this and other works of about the same time Erasmus showed a new boldness in commenting on the ills of Christian society—popes who in their warlike ambition imitated Caesar rather than Christ; princes who hauled whole nations into war to avenge a personal slight; and preachers who looked to their own interests by pronouncing the princes' wars just or by nurturing superstitious observances among the faithful. To remedy these evils Erasmus looked to education. In particular, the training of preachers should be based on "the philosophy of Christ" rather than on scholastic methods. Erasmus tried to show the way with his annotated text of the Greek New Testament and his edition of St. Jerome's *Opera omnia,* both of which appeared from the Froben press in 1516. These were the months in which Erasmus thought he saw "the world growing young again," and the full measure of his optimism is expressed in one of the prefatory writings to the New Testament: "If the Gospel were truly preached, the Christian people would be spared many wars."

The function of education

Erasmus' home base was now in Brabant, where he had influential friends at the Habsburg court of the Netherlands in Brussels, notably the grand chancellor, Jean Sauvage. Through Sauvage he was named honorary councillor to the 16-year-old archduke Charles, the future Charles V, and was commissioned to write *Institutio principis Christiani* (1516; *The Education of a Christian Prince*) and *Querela pacis* (1517; *The Complaint of Peace*). These works expressed Erasmus' own convictions, but they also did no harm to Sauvage's faction at court, which wanted to maintain peace with France. It was at this time too that he began his *Paraphrases* of the books of the New Testament, each one dedicated to a monarch or a prince of the church. He was accepted as a member of the theology faculty at nearby Louvain, and he also took keen interest in a newly founded Trilingual College, with endowed chairs in Latin, Greek, and Hebrew. *Ratio verae theologiae* (1518) provided the rationale for the new theological education based on the study of languages. Revision of his Greek New Testament, especially of the copious annotations, began almost as soon as the first edition appeared. Though Erasmus certainly made mistakes as a textual critic, in the history of scholarship he is a towering figure, intuiting philological principles that in some cases would not be formulated explicitly until 150 years after his death. But conservative theologians at Louvain and elsewhere, mostly ignorant of Greek, were not willing to abandon the interpretation of Scripture to upstart "grammarians," nor did the atmosphere at Louvain improve when the second edition of Erasmus' New Testament (1519) replaced the Vulgate with his own Latin translation.

The Protestant challenge. From the very beginning of the momentous events sparked by Martin Luther's challenge to papal authority, Erasmus' clerical foes blamed him for inspiring Luther, just as some of Luther's admirers in Germany found that he merely proclaimed boldly what Erasmus had been hinting. In fact, Luther's first letter to Erasmus (1516) showed an important disagreement over the interpretation of St. Paul, and in 1518 Erasmus privately instructed his printer, Froben, to stop printing works by Luther, lest the two causes be confused. As he read Luther's writings, at least those prior to *The Babylonian Captivity of the Church* (1520), Erasmus found much to admire, and he could even describe Luther, in a letter to Pope Leo X, as "a mighty trumpet of Gospel truth." Being of a suspicious nature, however, he also convinced himself that Luther's fiercest enemies were men who saw the study of languages as the root of heresy and thus wanted to be rid of both at once. Hence he tugged at the slender threads of his influence, vainly hoping to forestall a confrontation that could only be destructive to "good letters." When he quit Brabant for Basel (December 1521), he did so lest he be faced with a personal request from the Emperor to write a book against Luther, which he could not have refused.

Erasmus and Luther

Erasmus' belief in the unity of the church was fundamental, but, like the Hollanders and Brabanters with whom he was most at home, he recoiled from the cruel logic of religious persecution. He expressed his views indirectly through the *Colloquia,* which had started as schoolboy dialogues but now became a vehicle for commentary. For example, in the colloquy "Inquisitio de fide" (1522) a Catholic finds to his surprise that Lutherans accept all the dogmas of the faith, that is, the articles of the Apostles' Creed. The implication is that bitter disputes like those over papal infallibility or Luther's doctrine of predestination are differences over mere opinion, not over dogmas binding on all the faithful. For Erasmus the root of the schism was not theology but anticlericalism and lay resentment of the laws and "ceremonies" that the clergy made binding under pain of hell. As he wrote privately to the Netherlandish pope Adrian VI (1522–23), whom he had known at Louvain, there was still hope of reconciliation, if only the church would ease the burden; this could be accomplished, for instance, by granting the chalice to the laity and by permitting priests to marry: "At the sweet name of liberty all things will revive."

When Adrian VI was succeeded by Clement VII, Erasmus could no longer avoid "descending into the arena" of theological combat, though he promised the Swiss reformer Huldrych Zwingli that he would attack Luther in a way that would not please the "pharisees." *De libero arbitrio* (1524) defended the place of human free choice in the process of salvation and argued that the consensus of the church through the ages is authoritative in the interpretation of Scripture. In reply Luther wrote one of his most important theological works, *De servo arbitrio* (1525), to which Erasmus responded with a lengthy, two-part *Hyperaspistes* (1526–27). In this controversy Erasmus lets it be seen that he would like to claim more for free will than St. Paul and St. Augustine seem to allow.

Controversy over free will

The years in Basel (1522–29) were filled with polemics, some of them rather tiresome by comparison to the great debate with Luther. Irritated by Protestants who called him a traitor to the Gospel as well as by hyper-orthodox Catholic theologians who repeatedly denounced him, Erasmus showed the petty side of his own nature often enough. Although there is material in his apologetic writings that scholars have yet to exploit, there seems no doubt that on the whole he was better at satiric barbs, such as the

colloquy representing one young "Pseudo-Evangelical" of his acquaintance as thwacking people over the head with a Gospel book to gain converts. Meanwhile he kept at work on the Greek New Testament (there would be five editions in all), the *Paraphrases,* and his editions of the Church Fathers, including Cyprian, Hilary, and Origen. He also took time to chastise those humanists, mostly Italian, who from a "superstitious" zeal for linguistic purity refused to sully their Latin prose with nonclassical terms (*Ciceronianus,* 1528).

Final years. In 1529, when Protestant Basel banned Catholic worship altogether, Erasmus and some of his humanist friends moved to the Catholic university town of Freiburg im Breisgau. He refused an invitation to the Diet of Augsburg, where Philipp Melanchthon's Augsburg Confession was to initiate the first meaningful discussions between Lutheran and Catholic theologians. He nonetheless encouraged such discussion in *De sarcienda ecclesiae concordia* (1533), which suggested that differences on the crucial doctrine of justification might be reconciled by considering a *duplex justitia,* the meaning of which he did not elaborate. Having returned to Basel to see his manual on preaching (*Ecclesiastes,* 1535) through the press, he lingered on in a city he found congenial; it was there he died on July 12, 1536. Like the disciples of Voirier, he seems not to have asked for the last sacraments of the church. His last words were in Dutch: "*Lieve God*" ("dear God").

Influence and achievement. Always the scholar, Erasmus could see many sides of an issue. But his hesitations and studied ambiguities were appreciated less and less in the generations that followed his death, as men girded for combat, theological or otherwise, in the service of their beliefs. For a time, while peacemakers on both sides had an opportunity to pursue meaningful discussions between Catholics and Lutherans, some of Erasmus' practical suggestions and his moderate theological views were directly pertinent. Even after ecumenism dwindled to a mere wisp of possibility, there were a few men willing to make themselves heirs of Erasmus' lonely struggle for a middle ground, like Jacques-Auguste de Thou in France and Hugo Grotius in the Netherlands; significantly, both were strong supporters of state authority and hoped to limit the influence of the clergy of their respective established churches. This tradition was perhaps strongest in the Netherlands, where Dirck Volckertszoon Coornhert and others found support in Erasmus for their advocacy of limited toleration for religious dissenters. Meanwhile, however, the Council of Trent and the rise of Calvinism ensured that such views were generally of marginal influence. The Catholic *index expurgatorius* of 1571 contained a long list of suspect passages to be deleted from any future editions of Erasmus' writings, and those Protestant tendencies that bear some comparison to Erasmus' defense of free will—current among the Philippists in Germany and the Arminians in the Netherlands—were bested by defenders of a sterner orthodoxy. Even in the classroom, Erasmus' preference for putting students directly in contact with the classics gave way to the use of compendiums and manuals of humanist rhetoric and logic that resembled nothing so much as the scholastic curriculum of the past. Similarly, the bold and independent scholarly temper with which Erasmus approached the text of the New Testament was for a long time submerged by the exigencies of theological polemics.

Erasmus' reputation began to improve in the late 17th century, when the last of Europe's religious wars was fading into memory and scholars like Richard Simon and Jean Le Clercq (the editor of Erasmus' works) were once again taking a more critical approach to biblical texts. By Voltaire's time, in the 18th century, it was possible to imagine that the clever and rather skeptical Erasmus must have been a philosophe before his time, one whose professions of religious devotion and submission to church authority could be seen as convenient evasions. This view of Erasmus, curiously parallel to the strictures of his orthodox critics, was long influential. Only in the past several decades have scholars given due recognition to the fact that the goal of his work was a Christianity purified by a deeper knowledge of its historic roots. Yet it was not entirely wrong to compare Erasmus with those Enlightenment thinkers who, like Voltaire, defended individual liberty at every turn and had little good to say about the various corporate solidarities by which human society holds together. Some historians would now trace the enduring debate between these complementary aspects of Western thought as far back as the 12th century, and in this very broad sense Erasmus and Voltaire are on the same side of a divide, just as, for instance, Machiavelli and Rousseau are on the other. In a unique manner that fused his multiple identities—as Netherlander, Renaissance humanist, and pre-Tridentine Catholic—Erasmus helped to build what may be called the liberal tradition of European culture.

MAJOR WORKS

THEOLOGICAL WORKS: *Enchiridion militis Christiani* (1503; *The Manuell of the Christen Knyght,* trans. by W. Tyndale, 1533); *Annotationes in Novum Testamentum* (1516); *Paraphrases in Novum Testamentum* (1517; *Paraphrase of Erasmus upon the Newe Testamente,* 1548); *Ratio verae theologiae* (1519); *De libero arbitrio diatribe* (1524); *Hyperaspistes diatribae adversus servum arbitrium Martini Lutheri* (1526).

EDUCATIONAL AND OCCASIONAL WRITINGS: *Adagia* (1500; *Proverbs or Adagies,* 1539); *Moriae encomium* (1511; *The Praise of Folie,* 1549); *Institutio principis Christiani* (1516; *The Education of a Christian Prince,* 1936); *Querela pacis* (1517; *The Complaint of Peace,* 1559); *Colloquia* (1522–33); *Ciceronianus* (1528; *Ciceronianus: or, A Dialogue on the Best Style of Speaking,* 1908); *De pueris instituendis* (1529).

COLLECTIONS AND TRANSLATIONS: *Opera omnia, emendatiora et auctiora,* 10 vol. in 11, ed. by Jean Leclerq (1703–06, reprinted 1961–62), remains the most complete and authoritative among the early editions. *Opera omnia Desiderii Erasmi Roterodami: recognita et adnotatione critica instructa notisque illustrata* (1969–) is a modern critical and annotated edition by an international team of scholars, under the auspices of the Royal Academy of The Netherlands. The multivolume edition is arranged not chronologically but according to the canon laid down by Erasmus himself. *Opus epistolarum Des. Erasmi Roterdami,* ed. by P.S. Allen, 12 vol. (1906–58), is a standard edition of the correspondence of Erasmus, whose letters are indispensable for any understanding of his work. For translations, the ongoing series *Collected Works of Erasmus* (1974–), published by the University of Toronto Press, has set high standards for accuracy. Other notable translations include *The "Adages" of Erasmus: A Study with Translations,* by Margaret Mann Phillips (1964); and *The Colloquies of Erasmus,* trans. by Craig R. Thompson (1965).

BIBLIOGRAPHY

Life and intellectual development: PAUL MESTWERDT, *Die Anfänge des Erasmus: Humanismus und "devotio moderna"* (1917, reprinted 1971); JOHAN HUIZINGA, *Erasmus of Rotterdam* (1952; originally published in Dutch, 1924); AUGUSTIN RENAUDET, *Érasme et l'Italie* (1954); ROLAND H. BAINTON, *Erasmus of Christendom* (1969, reissued 1982); and JAMES D. TRACY, *Erasmus, the Growth of a Mind* (1972).

Humanist and educational writings: Analyses include WILLIAM HARRISON WOODWARD, *Desiderius Erasmus Concerning the Aim and Method of Education* (1904); OTTO SCHOTTENLOHER, *Erasmus im Ringen um die humanistische Bildungsform* (1933); and JACQUES CHOMARAT, *Grammaire et rhétorique chez Erasme,* 2 vol. (1981).

Theology and religious thought: ALFONS AUER, *Die vollkommene Frömmigkeit des Christen: nach dem Enchiridion militis Christiani des Erasmus von Rotterdam* (1954); C. AUGUSTIJN, *Erasmus en de reformatie* (1962); HARRY MCSORLEY, *Luther: Right or Wrong?: An Ecumenical-Theological Study of Luther's Major Work, The Bondage of the Will* (1968; originally published in German, 1967); JOHN B. PAYNE, *Erasmus: His Theology of the Sacraments* (1970); and GEORGES CHANTRAINE, *Erasme et Luther libre et serf arbitre: étude historique et théologique* (1981).

Scholarly work and views: JERRY H. BENTLEY, *Humanists and Holy Writ: New Testament Scholarship in the Renaissance* (1983); ERIKA RUMMEL, *Erasmus as a Translator of the Classics* (1985), and *Erasmus' Annotations on the New Testament: From Philologist to Theologian* (1986); MARCEL BATAILLON, *Érasme et l'Espagne: recherches sur l'histoire spirituelle du XVI^e siècle* (1937); ANDREAS FLITNER, *Erasmus im Urteil seiner Nachwelt* (1952); GUIDO KISCH, *Erasmus und die Jurisprudenz seiner Zeit* (1960); JAMES D. TRACY, *The Politics of Erasmus: A Pacifist Intellectual and His Political Milieu* (1978); and BRUCE MANSFIELD, *Phoenix of His Age: Interpretations of Erasmus c. 1550–1750* (1979).

(J.D.T.)

Ethics

How should we live? Shall we aim at happiness or at knowledge, virtue, or the creation of beautiful objects? If we choose happiness, will it be our own or the happiness of all? And what of the more particular questions that face us: Is it right to be dishonest in a good cause? Can we justify living in opulence while elsewhere in the world people are starving? If conscripted to fight in a war we do not support, should we disobey the law? What are our obligations to the other creatures with whom we share this planet and to the generations of humans who will come after us?

Ethics deals with such questions at all levels. Its subject consists of the fundamental issues of practical decision making, and its major concerns include the nature of ultimate value and the standards by which human actions can be judged right or wrong.

The terms ethics and morality are closely related. We now often refer to ethical judgments or ethical principles where it once would have been more common to speak of moral judgments or moral principles. These applications are an extension of the meaning of ethics. Strictly speaking, however, the term refers not to morality itself but to the field of study, or branch of inquiry, that has morality as its subject matter. In this sense, ethics is equivalent to moral philosophy.

Although ethics has always been viewed as a branch of philosophy, its all-embracing practical nature links it with many other areas of study, including anthropology, biology, economics, history, politics, sociology, and theology. Yet, ethics remains distinct from such disciplines because it is not a matter of factual knowledge in the way that the sciences and other branches of inquiry are. Rather, it has to do with determining the nature of normative theories and applying these sets of principles to practical moral problems.

This article is divided into the following sections:

The origins of ethics

MYTHICAL ACCOUNTS

When did ethics begin and how did it originate? If we are referring to ethics proper—*i.e.*, the systematic study of what we ought to do—it is clear that ethics can only have come into existence when human beings started to reflect on the best way to live. This reflective stage emerged long after human societies had developed some kind of morality, usually in the form of customary standards of right and wrong conduct. The process of reflection tended to arise from such customs, even if in the end it may have found them wanting. Accordingly, ethics began with the introduction of the first moral codes.

Virtually every human society has some form of myth to explain the origin of morality. In the Louvre in Paris there is a black Babylonian column with a relief showing the sun god Shamash presenting the code of laws to Hammurabi. The Old Testament account of God giving the Ten Commandments to Moses on Mt. Sinai might be considered another example. In Plato's *Protagoras* there is an avowedly mythical account of how Zeus took pity on the hapless humans, who, living in small groups and with inadequate teeth, weak claws, and lack of speed, were no match for the other beasts. To make up for these deficiencies, Zeus gave humans a moral sense and the capacity for law and justice, so that they could live in larger communities and cooperate with one another.

Notion of divine origin

That morality should be invested with all the mystery and power of divine origin is not surprising. Nothing else could provide such strong reasons for accepting the moral law. By attributing a divine origin to morality, the priesthood became its interpreter and guardian, and thereby secured for itself a power that it would not readily relinquish. This link between morality and religion has been so firmly forged that it is still sometimes asserted that there

can be no morality without religion. According to this view, ethics ceases to be an independent field of study. It becomes, instead, moral theology.

There is some difficulty, already known to Plato, with the view that morality was created by a divine power. In his dialogue *Euthyphro,* Plato considered the suggestion that it is divine approval that makes an action good. Plato pointed out that if this were the case, we could not say that the gods approve of the actions because the actions are good. Why then do the gods approve of these actions rather than others? Is their approval entirely arbitrary? Plato considered this impossible and so held that there must be some standards of right or wrong that are independent of the likes and dislikes of the gods. Modern philosophers have generally accepted Plato's argument because the alternative implies that if the gods had happened to approve of torturing children and to disapprove of helping one's neighbours, then torture would have been good and neighbourliness bad.

A modern theist might say that since God is good, he could not possibly approve of torturing children nor disapprove of helping neighbours. In saying this, however, the theist would have tacitly admitted that there is a standard of goodness that is independent of God. Without an independent standard, it would be pointless to say that God is good; this could only mean that God is approved of by God. It seems therefore that, even for those who believe in the existence of God, it is impossible to give a satisfactory account of the origin of morality in terms of a divine creation. We need a different account.

There are other possible connections between religion and morality. It has been said that even if good and evil exist independently of God or the gods, only divine revelation can reliably inform us about good and evil. An obvious problem with this view is that those who receive divine revelations, or who consider themselves qualified to interpret them, do not always agree on what is good and what is evil. Without an accepted criterion for the authenticity of a revelation or an interpretation, we are no better off, so far as reaching moral agreement is concerned, than we would be if we were to decide on good and evil ourselves with no assistance from religion.

Traditionally, a more important link between religion and ethics was that religious teachings were thought to provide a reason for doing what is right. In its crudest form, the reason was that those who obey the moral law will be rewarded by an eternity of bliss while everyone else roasts in hell. In more sophisticated versions, the motivation provided by religion was less blatantly self-seeking and more of an inspirational kind. Whether in its crude or sophisticated version, or something in between, religion does provide an answer to one of the great questions of ethics: Why should I do what is right? As will be seen in the course of this article, however, the answer provided by religion is by no means the only answer. It will be considered after the alternatives have been examined.

PREHUMAN ETHICS

Can we do better than the religious accounts of the origin of morality? Because, for obvious reasons, we have no historical record of a human society in the period before it had any standards of right and wrong, history cannot tell us the origins of morality. Nor is anthropology able to assist because all human societies studied have already had, except perhaps during the most extreme circumstances, their own form of morality. Fortunately there is another mode of inquiry open to us. Human beings are social animals. Living in a social group is a characteristic we share with many other animal species, including our closest relatives, the apes. Presumably, the common ancestor of humans and apes also lived in a social group, so that we were social beings before we were human beings. Here, then, in the social behaviour of nonhuman animals and in the evolutionary theory that explains such behaviour, we may find the origins of human morality.

Social life, even for nonhuman animals, requires constraints on behaviour. No group can stay together if its members make frequent, no-holds-barred attacks on one another. Social animals either refrain altogether from attacking other members of the social group, or, if an attack does take place, the ensuing struggle does not become a fight to the death—it is over when the weaker animal shows submissive behaviour. It is not difficult to see analogies here with human moral codes. The parallels, however, go much further than this. Like humans, social animals may behave in ways that benefit other members of the group at some cost or risk to themselves. Male baboons threaten predators and cover the rear as the troop retreats. Wolves and wild dogs bring meat back to members of the pack not present at the kill. Gibbons and chimpanzees with food will, in response to a gesture, share their food with others of the group. Dolphins support sick or injured animals, swimming under them for hours at a time and pushing them to the surface so they can breathe.

Apparent altruistic behaviour among nonhuman animals

It may be thought that the existence of such apparently altruistic behaviour is odd, for evolutionary theory states that those who do not struggle to survive and reproduce will be wiped out in the ruthless competition known as natural selection. Research in evolutionary theory applied to social behaviour, however, has shown that evolution need not be quite so ruthless after all. Some of this altruistic behaviour is explained by kin selection. The most obvious examples are those in which parents make sacrifices for their offspring. If wolves help their cubs to survive, it is more likely that genetic characteristics, including the characteristic of helping their own cubs, will spread through further generations of wolves.

Less obviously, the principle also holds for assistance to other close relatives, even if they are not descendants. A child shares 50 percent of the genes of each of its parents, but full siblings too, on the average, have 50 percent of their genes in common. Thus a tendency to sacrifice one's life for two or more of one's siblings could spread from one generation to the next. Between cousins, where only 12½ percent of the genes are shared, the sacrifice-to-benefit ratio would have to be correspondingly increased.

When apparent altruism is not between kin, it may be based on reciprocity. A monkey will present its back to another monkey, who will pick out parasites; after a time the roles will be reversed. Reciprocity may also be a factor in food sharing among unrelated animals. Such reciprocity will pay off, in evolutionary terms, as long as the costs of helping are less than the benefits of being helped and as long as animals will not gain in the long run by "cheating"—that is to say, by receiving favours without returning them. It would seem that the best way to ensure that those who cheat do not prosper is for animals to be able to recognize cheats and refuse them the benefits of cooperation the next time around. This is only possible among intelligent animals living in small, stable groups over a long period of time. Evidence supports this conclusion: reciprocal behaviour has been observed in birds and mammals, the clearest cases occurring among wolves, wild dogs, dolphins, monkeys, and apes.

In short, kin altruism and reciprocity do exist, at least in some nonhuman animals living in groups. Could these forms of behaviour be the basis of human ethics? There are good reasons for believing that they could. A surprising proportion of human morality can be derived from the twin bases of concern for kin and reciprocity. Kinship is a source of obligation in every human society. A mother's duty to look after her children seems so obvious that it scarcely needs to be mentioned. The duty of a married man to support and protect his family is almost equally as widespread. Duties to close relatives take priority over duties to more distant relatives, but in most societies even distant relatives are still treated better than strangers.

Concern for kin and reciprocity

If kinship is the most basic and universal tie between human beings, the bond of reciprocity is not far behind. It would be difficult to find a society that did not recognize, at least under some circumstances, an obligation to return favours. In many cultures this is taken to extraordinary lengths, and there are elaborate rituals of gift giving. Often the repayment has to be superior to the original gift, and this escalation can reach such extremes as to threaten the economic security of the donor. The huge "potlatch" feasts of certain American Indian tribes are a well-known example of this type of situation. Many Melanesian soci-

eties also place great importance on giving and receiving very substantial amounts of valuable items.

Many features of human morality could have grown out of simple reciprocal practices such as the mutual removal of parasites from awkward places. Suppose I want to have the lice in my hair picked out and I am willing in return to remove lice from someone else's hair. I must, however, choose my partner carefully. If I help everyone indiscriminately, I will find myself delousing others without getting my own lice removed. To avoid this, I must learn to distinguish between those who return favours and those who do not. In making this distinction, I am separating reciprocators and nonreciprocators and, in the process, developing crude notions of fairness and of cheating. I will strengthen my links with those who reciprocate, and bonds of friendship and loyalty, with a consequent sense of obligation to assist, will result.

This is not all. The reciprocators are likely to react in a hostile and angry way to those who do not reciprocate. Perhaps they will regard reciprocity as good and "right" and cheating as bad and "wrong." From here it is a small step to concluding that the worst of the nonreciprocators should be driven out of society or else punished in some way, so that they will not take advantage of others again. Thus a system of punishment and a notion of desert constitute the other side of reciprocal altruism.

Although kinship and reciprocity loom large in human morality, they do not cover the entire field. Typically, there are obligations to other members of the village, tribe, or nation even when these are strangers. There may also be a loyalty to the group as a whole that is distinct from loyalty to individual members of the group. It may be at this point that human culture intervenes. Each society has a clear interest in promoting devotion to the group and can be expected to develop cultural influences that exalt those who make sacrifices for the sake of the group and revile those who put their own interests too far ahead of the interests of the group. More tangible rewards and punishments may supplement the persuasive effect of social opinion. This is simply the start of a process of cultural development of moral codes.

Before considering the cultural variations in human morality and their significance for ethics, let us draw together this discussion of the origins of morality. Since we are dealing with a prehistoric period and morality leaves no fossils, any account of the origins of morality will necessarily remain to some extent speculative. It seems likely that morality is the gradual outgrowth of forms of altruism that exist in some social animals and that are the result of the usual evolutionary processes of natural selection. No myths are required to explain its existence.

ANTHROPOLOGY AND ETHICS

It is commonly believed that there are no ethical universals—*i.e.*, there is so much variation from one culture to another that no single principle or judgment is generally accepted. We have already seen that such is not the case. Of course, there are immense differences in the way in which the broad principles so far discussed are applied. The duty of children to their parents meant one thing in traditional Chinese society and means something quite different in contemporary Anglo-Saxon society. Yet, concern for kin and reciprocity to those who treat us well are considered good in virtually all human societies. Also, all societies have, for obvious reasons, some constraints on killing and wounding other members of the group.

Beyond that common ground, the variations in moral attitudes soon become more striking than the similarities. Man's fascination with such variations goes back a long way. The Greek historian Herodotus relates that Darius, king of Persia, once summoned Greeks before him and asked them how much he would have to pay them to eat their fathers' dead bodies. They refused to do it at any price. Then Darius brought in some Indians who by custom ate the bodies of their parents and asked them what would make them willing to burn their fathers' bodies. The Indians cried out that he should not mention so horrid an act. Herodotus drew the obvious moral: each nation thinks its own customs best.

Variations in morals were not systematically studied until the 19th century, when knowledge of the more remote parts of the globe began to increase. At the beginning of the 20th century, Edward Westermarck published *The Origin and Development of the Moral Ideas* (1906–08), two large volumes comparing differences among societies in such matters as the wrongness of killing (including killing in warfare, euthanasia, suicide, infanticide, abortion, human sacrifices, and duelling); whose duty it is to support children, the aged, or the poor; the forms of sexual relationship permitted; the status of women; the right to property and what constitutes theft; the holding of slaves; the duty to tell the truth; dietary restrictions; concern for nonhuman animals; duties to the dead; and duties to the gods. Westermarck had no difficulty in demonstrating tremendous diversity in all these issues. More recent, though less comprehensive, studies have confirmed that human societies can and do flourish while holding radically different views about all such matters.

As noted earlier, ethics itself is not primarily concerned with the description of moral systems in different societies. That task, which remains on the level of description, is one for anthropology or sociology. In contrast, ethics deals with the justification of moral principles. Nevertheless, ethics must take note of the variations in moral systems because it has often been claimed that this knowledge shows that morality is simply a matter of what is customary and is always relative to a particular society. According to this view, no ethical principles can be valid except in terms of the society in which they are held. Words such as good and bad just mean, it is claimed, "approved in my society" or "disapproved in my society," and so to search for an objective, or rationally justifiable, ethic is to search for what is in fact an illusion.

One way of replying to this position would be to stress the fact that there are some features common to virtually all human moralities. It might be thought that these common features must be the universally valid and objective core of morality. This argument would, however, invoive a fallacy. If the explanation for the common features is simply that they are advantageous in terms of evolutionary theory, that does not make them right. Evolution is a blind force incapable of conferring a moral imprimatur on human behaviour. It may be a fact that concern for kin is in accord with evolutionary theory, but to say that concern for kin is therefore right would be to attempt to deduce values from facts. As will be seen later, it is not possible to deduce values from facts in this manner. In any case, that something is universally approved does not make it right. If all human societies enslaved any tribe they could conquer, some freethinking moralists might still insist that slavery is wrong. They could not be said to be talking nonsense merely because they had few supporters. Similarly, then, universal support for principles of kinship and reciprocity cannot prove that these principles are in some way objectively justified.

This example illustrates the way in which ethics differs from a descriptive science. From the standpoint of ethics, whether human moral codes closely parallel one another or are extraordinarily diverse, the question of how an individual should act remains open. If you are thinking deeply about what you should do, your uncertainty will not be overcome by being told what your society thinks you should do in the circumstances in which you find yourself. Even if you are told that virtually all other human societies agree, you may choose not to go that way. If you are told that there is great variation among human societies over what people should do in your circumstances, you may wonder whether there can be any objective answer, but your dilemma has still not been resolved. In fact, this diversity does not rule out the possibility of an objective answer either: conceivably, most societies simply got it wrong. This, too, is something that will be taken up later in this article, for the possibility of an objective morality is one of the constant themes of ethics.

ANCIENT ETHICS

The first ethical precepts were certainly passed down by word of mouth by parents and elders, but as societies

learned to use the written word, they began to set down their ethical beliefs. These records constitute the first historical evidence of the origins of ethics.

The Middle East. The earliest surviving writings that might be taken as ethics textbooks are a series of lists of precepts to be learned by boys of the ruling class of Egypt, prepared some 3,000 years before the Christian Era. In most cases, they consist of shrewd advice on how to live happily, avoid unnecessary troubles, and advance one's career by cultivating the favour of superiors. There are, however, several passages that recommend more broadly based ideals of conduct, such as the following: Rulers should treat their people justly and judge impartially between their subjects. They should aim to make their people prosperous. Those who have bread are urged to share it with the hungry. Humble and lowly people must be treated with kindness. One should not laugh at the blind or at dwarfs.

Why then should one follow these precepts? Did the ancient Egyptians believe that one should do what is good for its own sake? The precepts frequently state that it will profit a man to act justly, much as we say that "honesty is the best policy." They also emphasize the importance of having a good name. Since these precepts are intended for the instruction of the ruling classes, however, we have to ask why helping the destitute should have contributed to an individual's good reputation among this class. To some degree the authors of the precepts must have thought that to make people prosperous and happy and to be kind to those who have least is not merely personally advantageous but good in itself.

The precepts are not works of ethics in the philosophical sense. No attempt is made to find any underlying principles of conduct that might provide a more systematic understanding of ethics. Justice, for example, is given a prominent place, but there is no elaboration of the notion of justice nor any discussion of how disagreements about what is just and unjust might be resolved. Furthermore, there is no probing of ethical dilemmas that may occur if the precepts should conflict with one another. The precepts are full of sound observations and practical wisdom, but they do not encourage theoretical speculation.

The same practical bent can be found in other early codes or lists of ethical injunctions. The great codification of Babylonian law by Hammurabi is often said to have been based on the principle of "an eye for an eye, a tooth for a tooth," as if this were some fundamental principle of justice, elaborated and applied to all cases. In fact, the code reflects no such consistent principle. It frequently prescribes the death penalty for offenses that do not themselves cause death—*e.g.,* for robbery or for accepting bribes. Moreover, even the eye-for-an-eye rule applies only if the eye of the original victim is that of a member of the patrician class; if it is the eye of a commoner, the punishment is a fine of a quantity of silver. Apparently such differences in punishment were not thought to require justification. At any rate, there are no surviving attempts to defend the principles of justice on which the code was based.

Code of Hammurabi

The Hebrew people were at different times captives of both the Egyptians and the Babylonians. It is therefore not surprising that the law of ancient Israel, which was put into its definitive form during the Babylonian Exile, shows the influence both of the ancient Egyptian precepts and of the Code of Hammurabi. The book of Exodus refers, for example, to the principle of "life for life, eye for eye, tooth for tooth." Hebrew law does not differentiate, as the Babylonian law does, between patricians and commoners, but it does stipulate that in several respects foreigners may be treated in ways that it is not permissible to treat fellow Hebrews; for instance, Hebrew slaves, but not others, had to be freed without ransom in the seventh year. Yet, in other respects Israeli law and morality developed the humane concern shown in the Egyptian precepts for the poor and unfortunate: hired servants must be paid promptly, because they rely on their wages to satisfy their pressing needs; slaves must be allowed to rest on the seventh day; widows, orphans, and the blind and deaf must not be wronged, and the poor man should not be refused a loan. There was even a tithe providing for an incipient welfare state. The spirit of this humane concern was summed up by the injunction to "love thy neighbour as thyself," a sweepingly generous form of the rule of reciprocity.

The famed Ten Commandments are thought to be a legacy of Semitic tribal law when important commands were taught, one for each finger, so that they could more easily be remembered. (Sets of five or 10 laws are common among preliterate civilizations.) The content of the Hebrew commandments differed from other laws of the region mainly in its emphasis on duties to God. In the more detailed laws laid down elsewhere, this emphasis continued with as much as half the legislation concerned with crimes against God and ceremonial and ritualistic matters, though there may be other explanations for some of these ostensibly religious requirements concerning the avoidance of certain foods and the need for ceremonial cleansings.

In addition to lengthy statements of the law, the surviving literature of ancient Israel includes both proverbs and the books of the prophets. The proverbs, like the precepts of the Egyptians, are brief statements without much concern for systematic presentation or overall coherence. They go further than the Egyptian precepts, however, in urging conduct that is just and upright and pleasing to God. There are correspondingly fewer references to what is needed for a successful career, although it is frequently stated that God rewards the just. In this connection the Book of Job is notable as an exploration of the problem raised for those who accept this motive for obeying the moral law: How are we to explain the fact that the best of people may suffer the worst misfortunes? The book offers no solution beyond faith in God, but the sharpened awareness of the problem it offers may have influenced some to adopt belief in reward and punishment in another realm as the only possible solution.

Hebrew proverbs and the books of the prophets

The literature of the prophets contains a good deal of social and ethical criticism, though more at the level of denunciation than discussion about what goodness really is or why there is so much wrongdoing. The Book of Isaiah is especially notable for its early portrayal of a utopia in which "the desert shall blossom as the rose . . . the wolf also shall dwell with the lamb . . . . They shall not hurt or destroy in all my holy mountain."

India. Unlike the ethical teaching of ancient Egypt and Babylon, Indian ethics was philosophical from the start. In the oldest of the Indian writings, the Vedas, ethics is an integral aspect of philosophical and religious speculation about the nature of reality. These writings date from about 1500 BC. They have been described as the oldest philosophical literature in the world, and what they say about how people ought to live may therefore be the first philosophical ethics.

The Vedas

The Vedas are, in a sense, hymns, but the gods to which they refer are not persons but manifestations of ultimate truth and reality. In the Vedic philosophy, the basic principle of the universe, the ultimate reality on which the cosmos exists, is the principle of *Ritam,* which is the word from which the Western notion of right is derived. There is thus a belief in a right moral order somehow built into the universe itself. Hence, truth and right are linked; to penetrate through illusion and understand the ultimate truth of human existence is to understand what is right. To be an enlightened one is to know what is real and to live rightly, for these are not two separate things but one and the same.

The ethic that is thus traced to the very essence of the universe is not without its detailed practical applications. These were based on four ideals, or proper goals, of life: prosperity, the satisfaction of desires, moral duty, and spiritual perfection—*i.e.,* liberation from a finite existence. From these ends follow certain virtues: honesty, rectitude, charity, nonviolence, modesty, and purity of heart. To be condemned, on the other hand, are falsehood, egoism, cruelty, adultery, theft, and injury to living things. Because the eternal moral law is part of the universe, to do what is praiseworthy is to act in harmony with the universe and accordingly will receive its proper reward; conversely, once the true nature of the self is understood, it becomes

apparent that those who do what is wrong are acting self-destructively.

The basic principles underwent considerable modification over the ensuing centuries, especially in the *Upaniṣads*, a body of philosophical literature dating from 800 BC. The Indian caste system, with its intricate laws about what members of each caste may or may not do, is accepted by the *Upaniṣads* as part of the proper order of the universe. Ethics itself, however, is not regarded as a matter of conformity to laws. Instead, the desire to be ethical is an inner desire. It is part of the quest for spiritual perfection, which in turn is elevated to the highest of the four goals of life.

During the following centuries the ethical philosophy of this early period gradually became a rigid and dogmatic system that provoked several reactions. One, which is uncharacteristic of Indian thought in general, was the Cārvāka, or materialist school, which mocked religious ceremonies, saying that they were invented by the Brahmans (the priestly caste) to ensure their livelihood. When the Brahmans defended animal sacrifices by claiming that the sacrificed beast goes straight to heaven, the members of the Cārvāka asked why the Brahmans did not kill their aged parents to hasten their arrival in heaven. Against the postulation of an eventual spiritual liberation, Cārvāka ethics urged each individual to seek his or her pleasure here and now.

The Cārvāka

Jainism, another reaction to the traditional Vedic outlook, went in exactly the opposite direction. The Jaina philosophy is based on spiritual liberation as the highest of all goals and nonviolence as the means to it. In true philosophical manner, the Jainas found in the principle of nonviolence a guide to all morality. First, apart from the obvious application to prohibiting violent acts to other humans, nonviolence is extended to all living things. The Jainas are vegetarian. They are often ridiculed by Westerners for the care they take to avoid injuring insects or other living things while walking or drinking water that may contain minute organisms; it is less well known that Jainas began to care for sick and injured animals thousands of years before animal shelters were thought of in Europe. The Jainas do not draw the distinction usually made in Western ethics between their responsibility for what they do and their responsibility for what they omit doing. Omitting to care for an injured animal would also be in their view a form of violence.

Jaina philosophy

Other moral duties are also derived from the notion of nonviolence. To tell someone a lie, for example, is regarded as inflicting a mental injury on that person. Stealing, of course, is another form of injury, but because of the absence of a distinction between acts and omissions, even the possession of wealth is seen as depriving the poor and hungry of the means to satisfy their wants. Thus nonviolence leads to a principle of nonpossession of property. Jaina priests were expected to be strict ascetics and to avoid sexual intercourse. Ordinary Jainas, however, followed a slightly less severe code, which was intended to give effect to the major forms of nonviolence while still being compatible with a normal life.

The other great ethical system to develop as a reaction to the ossified form of the old Vedic philosophy was Buddhism. The person who became known as the Buddha, which means the "enlightened one," was born about 563 BC, the son of a king. Until he was 29 years old, he lived the sheltered life of a typical prince, with every luxury he could desire. At that time, legend has it, he was jolted out of his idleness by the "Four Signs": he saw in rapid succession a very feeble old man, a hideous leper, a funeral, and a venerable ascetic monk. He began to think about old age, disease, and death, and decided to follow the way of the monk. For six years he led an ascetic life of renunciation, but finally, while meditating under a tree, he concluded that the solution was not withdrawal from the world, but rather a practical life of compassion for all.

Buddhism is often thought to be a religion, and indeed over the centuries it has adopted in many places the trappings of religion. This is an irony of history, however, because the Buddha himself was a strong critic of religion. He rejected the authority of the Vedas and refused to set up any alternative creed. He saw religious ceremonies as a waste of time and theological beliefs as mere superstition. He refused to discuss abstract metaphysical problems such as the immortality of the soul. The Buddha told his followers to think for themselves and take responsibility for their own future. In place of religious beliefs and religious ceremonies, the Buddha advocated a life devoted to universal compassion and brotherhood. Through such a life one might reach the ultimate goal, Nirvāṇa, a state in which all living things are free from pain and sorrow. There are similarities between this ethic of universal compassion and the ethics of the Jainas. Nevertheless, the Buddha was the first historical figure to develop such a boundless ethic.

The teachings of the Buddha

In keeping with his own previous experience, the Buddha proposed a "middle path" between self-indulgence and self-renunciation. In fact, it is not so much a path between these two extremes as one that draws together the benefits of both. Through living a life of compassion and love for all, a person achieves the liberation from selfish cravings sought by the ascetic and a serenity and satisfaction that are more fulfilling than anything obtained by indulgence in pleasure.

It is sometimes thought that because the Buddhist goal is Nirvāṇa, a state of freedom from pain and sorrow that can be reached by meditation, Buddhism teaches a withdrawal from the real world. Nirvāṇa, however, is not to be sought for oneself alone; it is regarded as a unity of the individual self with the universal self in which all things take part. In the Mahāyāna school of Buddhism, the aspirant for Enlightenment even takes a vow not to accept final release until everything that exists in the universe has attained Nirvāṇa.

The Buddha lived and taught in India, and so Buddhism is properly classified as an Indian ethical philosophy. Yet, Buddhism did not take hold in the land of its origin. Instead, it spread in different forms south into Sri Lanka and Southeast Asia, and north through Tibet to China, Korea, and Japan. In the process, Buddhism suffered the same fate as the Vedic philosophy against which it had rebelled: it became a religion, often rigid, with its own sects, ceremonies, and superstitions.

China. The two greatest moral philosophers of ancient China, Lao-tzu (flourished *c.* 6th century BC) and Confucius (551–479 BC), thought in very different ways. Lao-tzu is best known for his ideas about the Tao (literally "Way," the Supreme Principle). The Tao is based on the traditional Chinese virtues of simplicity and sincerity. To follow the Tao is not a matter of keeping to any set list of duties or prohibitions, but rather of living in a simple and honest manner, being true to oneself, and avoiding the distractions of ordinary living. Lao-tzu's classic book on the Tao, *Tao-te Ching*, consists only of aphorisms and isolated paragraphs, making it difficult to draw an intelligible system of ethics from it. Perhaps this is because Lao-tzu was a type of moral skeptic: he rejected both righteousness and benevolence, apparently because he saw them as imposed on individuals from without rather than coming from their own inner nature. Like the Buddha, Lao-tzu found the things prized by the world—rank, luxury, and glamour—to be empty, worthless values when compared with the ultimate value of the peaceful inner life. He also emphasized gentleness, calm, and nonviolence. Nearly 600 years before Jesus, he said: "It is the way of the Tao . . . to recompense injury with kindness." By returning good for good and also good for evil, Lao-tzu believed that all would become good; to return evil for evil would lead to chaos.

Lao-tzu

The lives of Lao-tzu and Confucius overlapped, and there is even an account of a meeting between them, which is said to have left the younger Confucius baffled. Confucius was the more down-to-earth thinker, absorbed in the practical task of social reform. When he was a provincial minister of justice, the province became renowned for the honesty of its people and their respect for the aged and their care for the poor. Probably because of its practical nature, the teachings of Confucius had a far greater influence on China than did those of the more withdrawn Lao-tzu.

Confucius did not organize his recommendations into

The teachings of Confucius

any coherent system. His teachings are offered in the form of sayings, aphorisms, and anecdotes, usually in reply to questions by disciples. They aim at guiding the audience in what is necessary to become a better person, a concept translated as "gentleman" or "the superior man." In opposition to the prevailing feudal ideal of the aristocratic lord, Confucius presented the superior man as one who is humane and thoughtful, motivated by the desire to do what is good rather than by personal profit. Beyond this, however, the concept is not discussed in any detail; it is only shown by diverse examples, some of them trite: "A superior man's life leads upwards . . . . The superior man is broad and fair; the inferior man takes sides and is petty . . . . A superior man shapes the good in man; he does not shape the bad in him."

One of the recorded sayings of Confucius is an answer to a request from a disciple for a single word that could serve as a guide to conduct for one's entire life. He replied: "Is not reciprocity such a word? What you do not want done to yourself, do not do to others." This rule is repeated several times in the Confucian literature and might be considered the supreme principle of Confucian ethics. Other duties are not, however, presented as derivative from this supreme principle, nor is the principle used to determine what is to be done when more specific duties—*e.g.*, duties to parents and duties to friends, both of which were given prominence in Confucian ethics—should clash.

Confucius did not explain why the superior man chose righteousness rather than personal profit. This question was taken up more than 100 years after his death by his follower Mencius, who asserted that humans are naturally inclined to do what is humane and right. Evil is not in human nature but is the result of poor upbringing or lack of education. But Confucius also had another distinguished follower, Hsün-tzu, who said that man's nature is to seek self-profit and to envy others. The rules of morality are designed to avoid the strife that would otherwise follow from this nature. The Confucian school was united in its ideal of the superior man but divided over whether such an ideal was to be obtained by allowing people to fulfill their natural desires or by educating them to control those desires.

Ancient Greece. Early Greece was the birthplace of Western philosophical ethics. The ideas of Socrates, Plato, and Aristotle, who flourished in the 5th and 4th centuries BC, will be discussed in the next section. The sudden blooming of philosophy during that period had its roots in the ethical thought of earlier centuries. In the poetic literature of the 7th and 6th centuries BC, there were, as in the early development of ethics in other cultures, ethical precepts but no real attempts to formulate a coherent overall ethical position. The Greeks were later to refer to the most prominent of these poets and early philosophers as the seven sages, and they are frequently quoted with respect by Plato and Aristotle. Knowledge of the thought of this period is limited, for often only fragments of original writings, along with later accounts of dubious accuracy, remain.

Pythagoras (*c.* 580–*c.* 500 BC), whose name is familiar because of the geometrical theorem that bears his name, is one such early Greek thinker about whom little is known. He appears to have written nothing at all, but he was the founder of a school of thought that touched on all aspects of life and that may have been a kind of philosophical and religious order. In ancient times the school was best known for its advocacy of vegetarianism, which, like that of the Jainas, was associated with the belief that after the death of the body, the human soul may take up residence in the body of an animal. Pythagoreans continued to espouse this view for many centuries, and classical passages in the works of such writers as Ovid and Porphyry opposing bloodshed and animal slaughter can be traced back to Pythagoras.

Pythagorean school of thought

Ironically, an important stimulus for the development of moral philosophy came from a group of teachers to whom the later Greek philosophers—Socrates, Plato, and Aristotle—were consistently hostile: the Sophists. This term was used in the 5th century to refer to a class of professional teachers of rhetoric and argument. The Sophists promised their pupils success in political debate and increased influence in the affairs of the city. They were accused of being mercenaries who taught their students to win arguments by fair means or foul. Aristotle said that Protagoras, perhaps the most famous of them, claimed to teach how "to make the weaker argument the stronger."

The Sophists

The Sophists, however, were more than mere teachers of rhetorical tricks. They saw their role as imparting the cultural and intellectual qualities necessary for success, and their involvement with argument about practical affairs led them to develop views about ethics. The recurrent theme in the views of the better known Sophists, such as Protagoras, Antiphon, and Thrasymachus, is that what is commonly called good and bad or just and unjust does not reflect any objective fact of nature but is rather a matter of social convention. It is to Protagoras that we owe the celebrated epigram summing up this theme, "Man is the measure of all things." Plato represents him as saying "Whatever things seem just and fine to each city, are just and fine for that city, so long as it thinks them so." Protagoras, like Herodotus, was an early social relativist, but he drew a moderate conclusion from his relativism. He argued that while the particular content of the moral rules may vary, there must be rules of some kind if life is to be tolerable. Thus Protagoras stated that the foundations of an ethical system needed nothing from the gods or from any special metaphysical realm beyond the ordinary world of the senses.

The Sophist Thrasymachus appears to have taken a more radical approach—if Plato's portrayal of his views is historically accurate. He explained that the concept of justice means nothing more than obedience to the laws of society, and, since these laws are made by the strongest political group in their own interests, justice represents nothing but the interests of the stronger. This position is often represented by the slogan "Might is right." Thrasymachus was probably not saying, however, that whatever the mightiest do really is right; he is more likely to have been denying that the distinction between right and wrong has any objective basis. Presumably he would then encourage his pupils to follow their own interests as best they could. He is thus an early representative of Skepticism about morals and perhaps of a form of egoism, the view that the rational thing to do is follow one's own interests.

It is not surprising that with ideas of this sort in circulation other thinkers should react by probing more deeply into ethics to see if the potentially destructive conclusions of some of the Sophists could be resisted. This reaction produced works that have served ever since as the cornerstone for the entire edifice of Western ethics.

Western ethics from Socrates to the 20th century

THE CLASSICAL PERIOD OF GREEK ETHICS

Socrates. "The unexamined life is not worth living," Socrates once observed. This thought typifies his questioning, philosophical approach to ethics. Socrates, who lived from about 470 BC until he was put to death in 399 BC, must be regarded as one of the greatest teachers of ethics. Yet, unlike other figures of comparable importance such as the Buddha or Confucius, he did not tell his audience how they should live. What Socrates taught was a method of inquiry. When the Sophists or their pupils boasted that they knew what justice, piety, temperance, or law was, Socrates would ask them to give an account of it and then show that the account offered was entirely inadequate. For instance, against the received wisdom that justice consists in keeping promises and paying debts, Socrates put forth the example of a person faced with an unusual situation: a friend from whom he borrowed a weapon has since become insane but wants the weapon back. Conventional morality gives no clear answer to this dilemma; therefore, the original definition of justice has to be reformulated. So the Socratic dialogue gets under way.

Because his method of inquiry threatened conventional beliefs, Socrates' enemies contrived to have him put to death on a charge of corrupting the youth of Athens. For those who saw adherence to the conventional moral

code as more desirable than the cultivation of an inquiring mind, the charge was appropriate. By conventional standards, Socrates was indeed corrupting the youth of Athens, but he himself saw the destruction of beliefs that could not stand up to criticism as a necessary preliminary to the search for true knowledge. Here, he differed from the Sophists with their moral relativism, for he thought that virtue is something that can be known and that the good person is the one who knows of what virtue, or justice, consists.

It is therefore not entirely accurate to see Socrates as contributing a method of inquiry but no positive views of his own. He believed in goodness as something that can be known, even though he did not himself profess to know it. He also thought that those who know what good is are in fact good. This latter belief seems peculiar today, because we make a sharp distinction between what is good and what is in a person's own interests. Accordingly, it does not seem surprising if people know what they ought morally to do but then proceed to do what is in their own interests instead. How to provide such people with reasons for doing what is right has been a major problem for Western ethics. Socrates did not see a problem here at all; in his view anyone who does not act well must simply be ignorant of the nature of goodness. Socrates could say this because in ancient Greece the distinction between goodness and self-interest was not made, or at least not in the clear-cut manner that it is today. The Greeks believed that virtue is good both for the individual and for the community. To be sure, they recognized that to live virtuously might not be the best way to prosper financially, but then they did not assume, as we are prone to do, that material wealth is a major factor in whether a person's life goes well or ill.

Plato. Socrates' greatest disciple, Plato (428/427–348/347 BC), accepted the key Socratic beliefs in the objectivity of goodness and in the link between knowing what is good and doing it. He also took over the Socratic method of conducting philosophy, developing the case for his own positions by exposing errors and confusions in the arguments of his opponents. He did this by writing his works as dialogues in which Socrates is portrayed as engaging in argument with others, usually Sophists. The early dialogues are generally accepted as reasonably accurate accounts of Socrates' views, but the later ones, written many years after the death of Socrates, use the latter as a mouthpiece for ideas and arguments that were Plato's rather than those of the historical Socrates.

In the most famous of Plato's dialogues, *Politeia* (*The Republic*), the imaginary Socrates is challenged by the following example: Suppose a person obtained the legendary ring of Gyges, which has the magical property of rendering the wearer invisible. Would that person still have any reason to behave justly? Behind this challenge lies the suggestion, made by the Sophists and still heard today, that the only reason for acting justly is that one cannot get away with acting unjustly. Plato's response to this challenge is a long argument developing a position that appears to go beyond anything the historical Socrates asserted. Plato maintained that true knowledge consists not in knowing particular things but in knowing something general that is common to all the particular cases. This is obviously derived from the way in which Socrates would press his opponents to go beyond merely describing particular good, or temperate, or just acts, and to give instead a general account of goodness, or temperance, or justice. The implication is that we do not know what goodness is unless we can give this general account. But the question then arises, what is it that we know when we know this general idea of goodness? Plato's answer seems to be that what we know is some general form or idea of goodness, which is shared by every particular thing that is good. Yet, if we are truly to be able to know this form or idea of goodness, it seems to follow that it must really exist. Plato accepts this implication. His theory of forms is the view that when we know what goodness is, we have knowledge of something that is the common element in virtue of which all good things are good and, at the same time, is some existing thing, the pure form of goodness.

It has been said that all of Western philosophy consists of footnotes to Plato. Certainly the central issue around which all of Western ethics has revolved can be traced back to the debate between the Sophists, on the one hand, with their claims that goodness and justice are relative to the customs of each society or, worse still, merely a disguise for the interests of the stronger, and, on the other, Plato's defense of the possibility of knowledge of an objective form or idea of goodness.

But even if we know what goodness or justice is, why should we act justly if we can profit by doing the opposite? This remaining part of the challenge posed by the legendary ring of Gyges is still to be answered, for even if we accept that goodness is objective, it does not follow that we all have sufficient reason to do what is good. Whether goodness leads to happiness is, as has been seen from the preceding discussion of early ethics in other cultures, a perennial topic for all who think about ethics. Plato's answer is that justice consists in harmony between the three elements of the soul: intellect, emotion, and desire. The unjust person lives in an unsatisfactory state of internal discord, trying always to overcome the discomfort of unsatisfied desire but never achieving anything better than the mere absence of want. The soul of the good person, on the other hand, is harmoniously ordered under the governance of reason, and the good person finds truly satisfying enjoyment in the pursuit of knowledge. Plato remarks that the highest pleasure, in fact, comes from intellectual speculation. He also gives an argument for the belief that the human soul is immortal; therefore, even if just individuals seem to be living in poverty or illness, the gods will not neglect them in the next life, and there they will have the greatest rewards of all. In summary, then, Plato asserts that we should act justly because in doing so we are "at one with ourselves and with the gods."

Plato's concept of justice

Today, this may seem like a strange account of justice and a farfetched view of what it takes to achieve human happiness. Plato does not recommend justice for its own sake, independently of any personal gains one might obtain from being a just person. This is characteristic of Greek ethics, with its refusal to recognize that there could be an irresolvable conflict between one's own interest and the good of the community. Not until Immanuel Kant, in the 18th century, does a philosopher forcefully assert the importance of doing what is right simply because it is right quite apart from self-interested motivation. To be sure, Plato must not be interpreted as holding that the motivation for each and every just act is some personal gain; on the contrary, the person who takes up justice will do what is just because it is just. Nevertheless, Plato accepts the assumption of his opponents that one could not recommend taking up justice in the first place unless doing so could be shown to be advantageous for oneself as well as for others.

In spite of the fact that many people now think differently about this connection between morality and self-interest, Plato's attempt to argue that those who are just are in the long run happier than those who are unjust has had an enormous influence on Western ethics. Like Plato's views on the objectivity of goodness, the claim that justice and personal happiness are linked has helped to frame the agenda for a debate that continues even today.

Aristotle. Plato founded a school of philosophy in Athens known as the Academy. Here Aristotle (384–322 BC), Plato's younger contemporary and only rival in terms of influence on the course of Western philosophy, came to study. Aristotle was often fiercely critical of Plato, and his writing is very different in style and content, but the time they spent together is reflected in a considerable amount of common ground. Thus Aristotle holds with Plato that the life of virtue is rewarding for the virtuous, as well as beneficial for the community. Aristotle also agrees that the highest and most satisfying form of human existence is that in which man exercises his rational faculties to the fullest extent. One major difference is that Aristotle does not accept Plato's theory of common essences, or universal ideas, existing independently of particular things. Thus he does not argue that the path to goodness is through knowledge of the universal form or idea of "the good."

The basis of Aristotle's ethics

Aristotle's ethics are based on his view of the universe. He saw it as a hierarchy in which everything has a function. The highest form of existence is the life of the rational being, and the function of lower beings is to serve this form of life. This led him to defend slavery—because he thought barbarians were less rational than Greeks and by nature suited to be "living tools"—and the killing of nonhuman animals for food or clothing. From this also came a view of human nature and an ethical theory derived from it. All living things, Aristotle held, have inherent potentialities and it is their nature to develop that potential to the full. This is the form of life properly suited to them and constitutes their goal. What, however, is the potentiality of human beings? For Aristotle this question turns out to be equivalent to asking what it is that is distinctive about human beings, and this, of course, is the capacity to reason. The ultimate goal of humans, therefore, is to develop their reasoning powers. When they do this, they are living well, in accordance with their true nature, and they will find this the most rewarding existence possible.

Aristotle thus ends up agreeing with Plato that the life of the intellect is the highest form of life; though having a greater sense of realism than Plato, he tempered this view with the suggestion that the best feasible life for humans must also have the goods of material prosperity and close friendships. Aristotle's argument for regarding the life of the intellect so highly, however, is different from that used by Plato; and the difference is significant because Aristotle committed a fallacy that has often been repeated. The fallacy is to assume that whatever capacity distinguishes humans from other beings is, for that very reason, the highest and best of their capacities. Perhaps the ability to reason is the best of our capacities, but we cannot be compelled to draw this conclusion from the fact that it is what is most distinctive of the human species.

A broader and still more pervasive fallacy underlies Aristotle's ethics. It is the idea that an investigation of human nature can reveal what we ought to do. For Aristotle, an examination of a knife would reveal that its distinctive quality is to cut, and from this we could conclude that a good knife would be a knife that cuts well. In the same way, an examination of human nature should reveal the distinctive quality of human beings, and from this we should be able to conclude what it is to be a good human being. This line of thought makes sense if we think, as Aristotle did, that the universe as a whole has a purpose and that we exist as part of such a goal-directed scheme of things, but its error becomes glaring once we reject this view and come to see our existence as the result of a blind process of evolution. Then we know that the standards of quality for knives are a result of the fact that knives are made with a specific purpose in mind and that a good knife is one that fills this purpose well. Human beings, however, were not made with any particular purpose in mind. Their nature is the result of random forces of natural selection and thus cannot, without further moral premises, determine how they ought to live.

Concept of the final end

It is to Aristotle that we owe the notion of the final end, or, as it was later called by medieval scholars, the *summum bonum*—the overall good for human beings. This can be found, Aristotle wrote, by asking why we do the things that we do. If we ask why we chop wood, the answer may be to build a fire; and if we ask why we build a fire, it may be to keep warm; but, if we ask why we keep warm, the answer is likely to be simply that it is pleasant to be warm and unpleasant to be cold. We can ask the same kind of questions about other activities; the answer always points, Aristotle thought, to what he called *eudaimonia*. This Greek word is usually translated as "happiness," but this is only accurate if we understand that term in its broadest sense to mean living a fulfilling, satisfying life. Happiness in the narrower sense of joy or pleasure would certainly be a concomitant of such a life, but it is not happiness in this narrower sense that is the goal.

In searching for the overall good, Aristotle separates what may be called instrumental goods from intrinsic goods. The former are good only because they lead to something else that is good; the latter are good in themselves. The distinction is neglected in the early lists of ethical precepts that were surveyed above, but it is of the first importance if a firmly grounded answer to questions about how one ought to live is to be obtained.

Aristotle is also responsible for much later thinking about the virtues one should cultivate. In his most important ethical treatise, the *Ethica Nicomachea* (*Nicomachean Ethics*), he sorts through the virtues as they were popularly understood in his day, specifying in each case what is truly virtuous and what is mistakenly thought to be so. Here, he uses the idea of the Golden Mean, which is essentially the same idea as the Buddha's middle path between self-indulgence and self-renunciation. Thus courage, for example, is the mean between two extremes: one can have a deficiency of it, which is cowardice, or one can have an excess of it, which is foolhardiness. The virtue of friendliness, to give another example, is the mean between obsequiousness and surliness.

Aristotle does not intend the idea of the mean to be applied mechanically in every instance: he says that in the case of the virtue of temperance, or self-restraint, it is easy to find the excess of self-indulgence in the physical pleasures, but the opposite error, insufficient concern for such pleasures, scarcely exists. (The Buddha, with his experience of the ascetic life of renunciation, would not have agreed.) This caution in the application of the idea is just as well, for while it may be a useful device for moral education, the notion of a mean cannot help us to discover new truths about virtue. We can only arrive at the mean if we already have a notion as to what is an excess and what is a defect of the trait in question, but this is not something to be discovered by a morally neutral inspection of the trait itself. We need a prior conception of the virtue in order to decide what is excessive and what is defective. To attempt to use the doctrine of the mean to define the particular virtues would be to travel in a circle.

Aristotle's list of the virtues differs from later Christian lists. Courage, temperance, and liberality are common to both periods, but Aristotle also includes a virtue that literally means "greatness of soul." This is the characteristic of holding a high opinion of oneself. The corresponding vice of excess is unjustified vanity, but the vice of deficiency is humility, which for Christians is a virtue.

Aristotle's discussion of the virtue of justice has been the starting point for almost all Western accounts. He distinguishes between justice in the distribution of wealth or other goods and justice in reparation, as, for example, in punishing someone for a wrong he has done. The key element of justice, according to Aristotle, is treating like cases alike—an idea that has set later thinkers the task of working out which similarities (need, desert, talent) are relevant. As with the notion of virtue as a mean, Aristotle's conception of justice provides a framework that needs to be filled in before it can be put to use.

Theoretical and practical reason

Aristotle distinguished between theoretical and practical wisdom. His concept of practical wisdom is significant, for it goes beyond merely choosing the means best suited to whatever ends or goals one may have. The practically wise person also has the right ends. This implies that one's ends are not purely a matter of brute desires or feelings; the right ends are something that can be known. It also gives rise to the problem that faced Socrates: How is it that people can know the difference between good and bad and still choose what is bad? As noted earlier, Socrates simply denied that this could happen, saying that those who did not choose the good must, appearances notwithstanding, be ignorant of what it is. Aristotle said that this view of Socrates was "plainly at variance with the observed facts" and, instead, offered a detailed account of the ways in which one can possess knowledge and yet not act on it because of lack of control or weakness of will.

LATER GREEK AND ROMAN ETHICS

In ethics, as in many other fields, the later Greek and Roman periods do not display the same penetrating insight as the Classic period of 5th- and 4th-century Greek civilization. Nevertheless, the two dominant schools of thought, Stoicism and Epicureanism, represent important approaches to the question of how one ought to live.

The Stoics. Stoicism had its origins in the views of

Socrates and Plato, as modified by Zeno and then by Chrysippus in the 3rd century BC. It gradually gained influence in Rome, chiefly through the teachings of Cicero (106–43 BC) and then later in the 1st century AD through those of Seneca. Remarkably, its chief proponents include both a slave, Epictetus, and an emperor, Marcus Aurelius. This is a fine illustration of the Stoic message that what is important is the pursuit of wisdom and virtue, a pursuit that is open to all human beings owing to their common capacity for reason and that can be carried out no matter what the external circumstances of their lives.

Today, the word stoic conjures up one who remains unmoved by the sorrows and afflictions that distress the rest of humanity. This is an accurate representation of a stoic ideal, but it must be placed in the context of a systematic approach to life. Plato held that human passions and physical desires are in need of regulation by reason (see above). The Stoics went further: they rejected passions altogether as a basis for deciding what is good or bad. Physical desires cannot simply be abolished, but when we become wise we appreciate the difference between wanting something and judging it to be good. Our desires make us want something, but only our reason can judge the goodness of what is wanted. If we are wise, we will identify with our reason, not with our desires; hence, we will not place our hopes on the attainment of our physical desires nor our anxieties on our failure to attain them. Wise Stoics will feel physical pain as others do, but in their minds they will know that physical pain leaves the true reasoning self untouched. The only thing that is truly good is to live in a state of wisdom and virtue. In aiming at such a life, we are not subject to the same play of fortune that afflicts us when we aim at physical pleasure or material wealth, for wisdom and virtue are matters of the intellect and under our own control. Moreover, if matters become too grim, there is always a way of ending the pain of the physical world. The Stoics were not reluctant to counsel suicide as a means of avoiding otherwise inescapable pain.

Rejection of passion in making moral judgments

Perhaps the most important legacy of Stoicism, however, is its conviction that all human beings share the capacity to reason. This led the Stoics to a fundamental sense of equality, which went beyond the limited Greek conception of equal citizenship. Thus Seneca claimed that the wise man will esteem the community of rational beings far above any particular community in which the accident of birth has placed him, and Marcus Aurelius said that common reason makes all individuals fellow citizens. The belief that human reasoning capacities are common to all was also important, because from it the Stoics drew the implication that there is a universal moral law, which all people are capable of appreciating. The Stoics thus strengthened the tradition that sees the universality of reason as the basis on which ethical relativism is to be rejected.

The Epicureans. While the modern use of the term stoic accurately represents at least a part of the Stoic philosophy, anyone taking the present-day meaning of epicure as a guide to the philosophy of Epicurus (341–270 BC) would go astray. True, the Epicureans regarded pleasure as the sole ultimate good and pain as the sole evil; and they did regard the more refined pleasures as superior, simply in terms of the quantity and durability of the pleasure they provided, to the coarser pleasures. To portray them as searching for these more refined pleasures by dining at the best restaurants and drinking the finest wines, however, is the reverse of the truth. By refined pleasures, Epicurus meant pleasures of the mind, as opposed to the coarse pleasures of the body. He taught that the highest pleasure obtainable is the pleasure of tranquillity, which is to be obtained by the removal of unsatisfied wants. The way to do this is to eliminate all but the simplest wants; these are then easily satisfied even by those who are not wealthy.

Epicurus developed his position systematically. To determine whether something is good, he would ask if it increased pleasure or reduced pain. If it did, it was good as a means; if it did not, it was not good at all. Thus justice was good but merely as an expedient arrangement to prevent mutual harm. Why not then commit injustice when we can get away with it? Only because, Epicurus says, the perpetual dread of discovery will cause painful anxiety.

The fundamentals of Epicurean ethics

Epicurus also exalted friendship, and the Epicureans were famous for the warmth of their personal relationships; but, again, they proclaimed that friendship is good only because of its tendency to create pleasure.

Both Stoic and Epicurean ethics can be seen as precursors of later trends in Western ethics: the Stoics of the modern belief in equality and the Epicureans of a Utilitarian ethic based on pleasure. The development of these ethical positions, however, was dramatically affected by the spreading from the East of a new religion that had its roots in a Jewish conception of ethics as obedience to a divine authority. With the conversion of Emperor Constantine I to Christianity by AD 313, the older schools of philosophy lost their sway over the thinking of the Roman Empire.

CHRISTIAN ETHICS FROM THE NEW TESTAMENT TO THE SCHOLASTICS

Ethics in the New Testament. Matthew reports Jesus as having said, in the Sermon on the Mount, that he came not to destroy the law of the prophets but to fulfill it. Indeed, when Jesus is regarded as a teacher of ethics, it is clear that he was more a reformer of the Hebrew tradition than a radical innovator. The Hebrew tradition had a tendency to place great emphasis on compliance with the letter of the law; the Gospel accounts of Jesus portray him as preaching against this "righteousness of the scribes and Pharisees," championing the spirit rather than the letter of the law. This spirit he characterized as one of love, for God and for one's neighbour. But since he was not proposing that the old teachings be discarded, he saw no need to develop a comprehensive ethical system. Christianity thus never really broke with the Jewish conception of morality as a matter of divine law to be discovered by reading and interpreting the word of God as revealed in the Scriptures.

This conception of morality had important consequences for the future development of Western ethics. The Greeks and Romans, and indeed thinkers such as Confucius too, did not have the Western conception of a distinctively moral realm of conduct. For them, everything that one did was a matter of practical reasoning, in which one could do well or poorly. In the more legalistic Judeo-Christian view, however, it is one thing to lack practical wisdom in, say, household budgeting, and a quite different and much more serious matter to fall short of what the moral law requires. This distinction between the moral and the nonmoral realms now affects every question in Western ethics, including the very way the questions themselves are framed.

Another consequence of the retention of the basically legalistic stance of Jewish ethics was that from the beginning Christian ethics had to deal with the question of how to judge the person who breaks the law from good motives or keeps it from bad motives. The latter half of this question was particularly acute because the Gospels describe Jesus as repeatedly warning of a coming resurrection of the dead at which time all would be judged and punished or rewarded according to their sins and virtues in this life. The punishments and rewards were weighty enough to motivate anyone who took this message seriously; and it was given added emphasis by the fact that it was not going to be long in coming. (Jesus said that it would take place during the lifetime of some of those listening to him.) This is, therefore, an ethic that invokes external sanctions as a reason for doing what is right, in contrast to Plato or Aristotle for whom happiness is an internal element of a virtuous life. At the same time, it is an ethic that places love above mere literal compliance with the law. These two aspects do not sit easily together. Can one love God and neighbour in order to be rewarded with eternal happiness in another life?

Basic principles of the Christian ethic

The fact that Jesus and Paul, too, believed in the imminence of the Second Coming led them to suggest ways of living that were scarcely feasible on any other assumption: taking no thought for the morrow; turning the other cheek; and giving away all one has. Even Paul's preference for celibacy rather than marriage and his grudging acceptance of the latter on the basis that "It is better to marry than to burn" makes some sense once we grasp that he was proposing ethical standards for what he thought would be

the last generation on earth. When the expected event did not occur and Christianity became the official religion of the vast and embattled Roman Empire, Christian leaders were faced with the awkward task of reinterpreting these injunctions in a manner more suited for a continuing society.

The new Christian ethical standards did lead to some changes in Roman morality. Perhaps the most vital was a new sense of the equal moral status of all human beings. As previously noted, the Stoics had been the first to elaborate this conception, grounding equality on the common capacity to reason. For Christians, humans are equal because they are all potentially immortal and equally precious in the sight of God. This caused Christians to condemn a wide variety of practices that had been accepted by both Greek and Roman moralists. Many of these related to the taking of innocent human life: from the earliest days Christian leaders condemned abortion, infanticide, and suicide. Even killing in war was at first regarded as wrong, and soldiers converted to Christianity had refused to continue to bear arms. Once the empire became Christian, however, this was one of the inconvenient ideas that had to yield. In spite of what Jesus had said about turning the other cheek, the church leaders declared that killing in a "just war" was not a sin. The Christian condemnation of killing in gladiatorial games, on the other hand, had a more permanent effect. Finally, but perhaps most importantly, while Christian emperors continued to uphold the legality of slavery, the Christian church accepted slaves as equals, admitted them to its ceremonies, and regarded the granting of freedom to slaves as a virtuous, if not obligatory, act. This moral pressure led over several hundred years to the gradual disappearance of slavery in Europe.

The Christian contribution to improving the position of slaves can also be linked with the distinctively Christian list of virtues. Some of the virtues described by Aristotle, as, for example, greatness of soul, are quite contrary in spirit to Christian virtues such as humility. In general, it can be said that the Greeks and Romans prized independence, self-reliance, magnanimity, and worldly success. By contrast, Christians saw virtue in meekness, obedience, patience, and resignation. As the Greeks and Romans conceived virtue, a virtuous slave was almost a contradiction in terms, but for Christians there was nothing in the state of slavery that was incompatible with the highest moral character.

Augustine. Christianity began with a set of scriptures incorporating many ethical injunctions but with no ethical philosophy. The first serious attempt to provide such a philosophy was made by St. Augustine of Hippo (354–430). Augustine was acquainted with a version of Plato's philosophy, and he developed the Platonic idea of the rational soul into a Christian view wherein humans are essentially souls, using their bodies as means to achieve their spiritual ends. The ultimate object remains happiness, as in Greek ethics, but Augustine saw happiness as consisting in a union of the soul with God after the body has died. It was through Augustine, therefore, that Christianity received the Platonic theme of the relative inferiority of bodily pleasures. There was, to be sure, a fundamental difference: whereas Plato saw this inferiority in terms of a comparison with the pleasures of philosophical contemplation in this world, Christians compared them unfavourably with the pleasures of spiritual existence in the next world. Moreover, Christians came to see bodily pleasures not merely as inferior but also as a positive threat to the achievement of spiritual bliss.

It was also important that Augustine could not accept the view, common to so many Greek and Roman philosophers, that philosophical reasoning was the path to wisdom and happiness. For a Christian, of course, the path had to be through love of God and faith in Jesus as the Saviour. The result was to be, for many centuries, a rejection of the use of unfettered reasoning powers in ethics.

Augustine was aware of the tension caused by the dual Christian motivations of love of God and neighbour, on the one hand, and reward and punishment in the afterlife, on the other. He came down firmly on the side of love, insisting that those who keep the moral law through fear of punishment are not really keeping it at all. But it is not ordinary human love, either, that suffices as a motivation for true Christian living. Augustine believed all men bear the burden of Adam's original sin, and so are incapable of redeeming themselves by their own efforts. Only the unmerited grace of God makes possible obedience to the "first greatest commandment" of loving God, and without such, one cannot fulfill the moral law. This view made a clear-cut distinction between Christians and pagan moralists, no matter how humble and pure the latter might be; only the former could be saved because only they could receive the blessing of divine grace. But this gain, as Augustine saw it, was purchased at the cost of denying that man is free to choose good or evil. Only Adam had this choice: he chose for all humanity, and he chose evil.

Aquinas and the moral philosophy of the Scholastics. At this point we may pass over more than 800 years in silence, for there were no major developments in ethics in the West until the rise of Scholasticism in the 12th and 13th centuries. Among the first of the significant works written during this time was a treatise on ethics by the French philosopher and theologian Peter Abelard (1079–1142). His importance in ethical theory lies in his emphasis on intentions. Abelard maintained, for example, that the sin of sexual wrongdoing consists not in the act of illicit sexual intercourse nor even in the desire for it, but in mentally consenting to that desire. In this he was far more modern than Augustine, with his doctrine of grace, and also more thoughtful than those who even today assert that the mere desire for what is wrong is as wrong as the act itself. Abelard saw that there is a problem in holding anyone morally responsible for the existence of mere physical desires. His ingenious solution was taken up by later medieval writers, and traces of it can still be found in modern discussions of moral responsibility.

Aristotle's ethical writings were not known to scholars in western Europe during Abelard's time. Latin translations became available only in the first half of the 13th century, and the rediscovery of Aristotle dominated later medieval philosophy. Nowhere is his influence more marked than in the thought of St. Thomas Aquinas (1225–74), often regarded as the greatest of the Scholastic philosophers and undoubtedly the most influential, since his teachings became the semiofficial philosophy of the Roman Catholic Church. Such is the respect in which Aquinas held Aristotle that he referred to him simply as The Philosopher, and it is not too far from the truth to say that the chief aim of Aquinas' work was to reconcile Aristotle's views with Christian doctrine.

Reconciliation of Aristotelian views with Christian doctrine

Aquinas took from Aristotle the notion of a final end, or *summum bonum,* at which all action is ultimately directed; and, like Aristotle, he saw this end as necessarily linked with happiness. This conception was Christianized, however, by the idea that happiness is to be found in the love of God. Thus a person seeks to know God but cannot fully succeed in this in life on earth. The reward of heaven, where one can know God, is available only to those who merit it, though even then it is given by God's grace rather than obtained by right. Short of heaven, a person can experience only a more limited form of happiness to be gained through a life of virtue and friendship, much as Aristotle had recommended.

The blend of Aristotle's teachings and Christianity is also evident in Aquinas' views about right and wrong, and how we come to know the difference between them. Aquinas is often described as advocating a "natural law" ethic, but this term is easily misunderstood. The natural law to which Aquinas referred does not require a legislator any more than do the laws of nature that govern the motions of the planets. An even more common mistake is to imagine that this conception of natural law relies on contrasting what is natural with what is artificial. Aquinas' theory of the basis of right and wrong developed rather as an alternative to the view that morality is determined simply by the arbitrary will of God. Instead of conceiving of right and wrong in this manner as something fundamentally unrelated to human goals and purposes, Aquinas saw morality as deriving from human nature and the activities that are objectively suited to it.

It is a consequence of this natural law ethic that the difference between right and wrong can be appreciated by the use of reason and reflection on experience. Christian revelation may supplement this knowledge in some respects, but even such pagan philosophers as Aristotle could understand the essentials of virtuous living. We are, however, likely to err when we apply these general principles to the particular cases that confront us in everyday life. Corrupt customs and poor moral education may obscure the messages of natural reason. Hence, societies must enact laws of their own to supplement natural law and, where necessary, to coerce those who, because of their own imperfections, are liable to do what is wrong and socially destructive.

It follows, too, that virtue and human flourishing are linked. When we do what is right, we do what is objectively suited to our true nature. Thus the promise of heaven is no mere external sanction, rewarding actions that would otherwise be indifferent to us or even against our best interests. On the contrary, Aquinas wrote that "God is not offended by us except by what we do against our own good." Reward and punishment in the afterlife reinforce a moral law that all humans, Christian or pagan, have adequate prior reasons for following.

In arguing for his views, Aquinas was always concerned to show that he had the authority of the Scriptures or the Church Fathers on his side, but the substance of his ethical system is to a remarkable degree based on reason rather than revelation. This is strong testimony to the power of Aristotle's example. Nonetheless, Aquinas absorbed the weaknesses as well as the strengths of the Aristotelian system. His attempt to base right and wrong on human nature, in particular, invites the objection that we cannot presuppose our nature to be good. Aquinas might reply that it is good because God made it so, but this merely shifts back one step the issue of the basis of good and bad: Did God make it good in accordance with some independent standard of goodness, or would any human nature made by God be good? If we give the former answer, we need an account of the independent standard of goodness. Because this cannot—if we are to avoid circular argument—be based on human nature, it is not clear what account Aquinas could offer. If we maintain, however, that any human nature made by God would be good, we must accept that if God had made our nature such that we flourish and achieve happiness by torturing the weak and helpless among us, that would have been what we should do in order to live virtuously.

Something resembling this second option—but without the intermediate step of an appeal to human nature—was the position taken by the last of the great Scholastic philosophers, William of Ockham (*c.* 1285–1349?). Ockham boldly broke with much that had been taken for granted by his immediate predecessors. Fundamental to this was his rejection of the central Aristotelian idea that all things have a final end, or goal, toward which they naturally tend. He, therefore, also spurned Aquinas' attempt to base morality on human nature, and with it the idea that happiness is man's goal and closely linked with goodness. This led him to a position in stark contrast to almost all previous Western ethics. Ockham denied all standards of good and evil that are independent of God's will. What God wills is good; what God condemns is evil. That is all there is to say about the matter. This position is sometimes called a divine approbation theory, because it defines "good" as whatever is approved by God. As indicated earlier, when discussing attempts to link morality with religion, it follows from such a position that it is meaningless to describe God himself as good. It also follows that if God had willed us to torture children, it would be good to do so. As for the actual content of God's will, according to Ockham, that is not a subject for philosophy but rather a matter for revelation and faith.

Theory of divine approbation

The rigour and consistency of Ockham's philosophy made it for a time one of the leading schools of Scholastic thought, but eventually it was the philosophy of Aquinas that prevailed in the Roman Catholic Church. After the Reformation, however, Ockham's view exerted influence on Protestant theologians. Meanwhile, it hastened the decline of Scholastic moral philosophy because it effectively removed ethics from the sphere of reason.

RENAISSANCE AND REFORMATION

The revival of Classical learning and culture that began in 15th-century Italy and then slowly spread throughout Europe did not give immediate birth to any major new ethical theories. Its significance for ethics lies, rather, in a change of focus. For the first time since the conversion of the Roman Empire to Christianity, man, not God, became the chief object of interest, and the theme was not religion but humanism—the powers, freedom, and accomplishments of human beings. This does not mean that there was a sudden conversion to atheism. Renaissance thinkers remained Christian and still considered human beings as somehow midway between the beasts and the angels. Yet, even this middle position meant that humans were special. It meant, too, a new conception of human dignity and of the importance of the individual.

Machiavelli. Although the Renaissance did not produce any outstanding moral philosophers, there is one writer whose work is of some importance in the history of ethics: the Italian author and statesman Niccolò Machiavelli. His book *Il principe* (1513; *The Prince*) offered advice to rulers as to what they must do to achieve their aims and secure their power. Its significance for ethics lies precisely in the fact that Machiavelli's advice ignores the usual ethical rules: "It is necessary for a prince, who wishes to maintain himself, to learn how not to be good, and to use this knowledge and not use it, according to the necessities of the case." There had not been so frank a rejection of morality since the Greek Sophists. So startling is the cynicism of Machiavelli's advice that it has been suggested that *Il principe* was an attempt to satirize the conduct of the princely rulers of Renaissance Italy. It may be more accurate, however, to view Machiavelli as an early political scientist, concerned only with setting out what human beings are like and how power is maintained, with no intention of passing moral judgment on the state of affairs described. In any case, *Il principe* gained instant notoriety, and Machiavelli's name became synonymous with political cynicism and deviousness. In spite of the chorus of condemnation, the work has led to a sharper appreciation of the difference between the lofty ethical systems of the philosophers and the practical realities of political life.

The first Protestants. It was left to the 17th-century English philosopher and political theorist Thomas Hobbes to take up the challenge of constructing an ethical system on the basis of so unflattering a view of human nature (see below). Between Machiavelli and Hobbes, however, there occurred the traumatic breakup of Western Christianity known as the Reformation. Reacting against the worldly immorality apparent in the Renaissance church, Martin Luther, John Calvin, and other leaders of the new Protestantism sought to return to the pure early Christianity of the Scriptures, especially the teachings of Paul, and of the Church Fathers, with Augustine foremost among them. They were contemptuous of Aristotle (Luther called him a "buffoon") and of non-Christian philosophers in general. Luther's standard of right and wrong was what God commands. Like William of Ockham, Luther insisted that the commands of God cannot be justified by any independent standard of goodness: good simply means what God commands. Luther did not believe these commands would be designed to satisfy human desires because he was convinced that desires are totally corrupt. In fact, he thought that human nature was totally corrupt. In any case, Luther insisted that one does not earn salvation by good works: one is justified by faith in Christ and receives salvation through divine grace.

Luther's standard of right and wrong

It is apparent that if these premises are accepted, there is little scope for human reason in ethics. As a result, no moral philosophy has ever had the kind of close association with any Protestant church that, say, the philosophy of Aquinas has had with Roman Catholicism. Yet, because Protestants emphasized the capacity of the individual to read and understand the Gospels without obtaining the authoritative interpretation of the church, the ultimate outcome of the Reformation was a greater freedom to

read and write independently of the church hierarchy. This made possible a new era of ethical thought.

From this time, too, distinctively national traditions of moral philosophy began to emerge; the British tradition, in particular, developed largely independently of ethics on the Continent. Accordingly, the present discussion will follow this tradition through the 19th century before returning to consider the different line of development in continental Europe.

THE BRITISH TRADITION: FROM HOBBES TO THE UTILITARIANS

Hobbes. Thomas Hobbes (1588–1679) is an outstanding example of the independence of mind that became possible in Protestant countries after the Reformation. God does, to be sure, play an honourable role in Hobbes's philosophy, but it is a dispensable role. The philosophical edifice stands on its own foundations; God merely crowns the apex. Hobbes was the equal of the Greek philosophers in his readiness to develop an ethical position based only on the facts of human nature and the circumstances in which humans live; and he surpassed even Plato and Aristotle in the extent to which he sought to do this by systematic deduction from clearly set out premises.

Hobbes started with a severe view of human nature: all of man's voluntary acts are aimed at self-pleasure or self-preservation. This position is known as psychological hedonism, because it asserts that the fundamental psychological motivation is the desire for pleasure. Like later psychological hedonists, Hobbes was confronted with the objection that people often seem to act altruistically. There is a story that Hobbes was seen giving alms to a beggar outside St. Paul's Cathedral. A clergyman sought to score a point by asking Hobbes if he would have given the money, had Christ not urged giving to the poor. Hobbes replied that he gave the money because it pleased him to see the poor man pleased. The reply reveals the dilemma that always faces those who propose startling new explanations for all human actions: either the theory is flagrantly at odds with how people really behave or else it must be broadened to such an extent that it loses much of what made it so shocking in the first place.

The psychological hedonism of Hobbes

Hobbes's account of "good" is equally devoid of religious or metaphysical premises. He defined good as "any object of desire," and insisted that the term must be used in relation to a person—nothing is simply good of itself independently of the person who desires it. Hobbes may therefore be considered a subjectivist. If one were to say, for example, of the incident just described, "What Hobbes did was good," this statement would not be objectively true or false. It would be good for the poor man, and, if Hobbes's reply was accurate, it would also be good for Hobbes. But if a second poor person, for instance, was jealous of the success of the first, that person could quite properly say that what Hobbes did was bad.

Remarkably, this unpromising picture of self-interested individuals who have no notion of good apart from their own desires serves as the foundation of Hobbes's account of justice and morality in his masterpiece, *Leviathan* (1651). Starting with the premises that humans are self-interested and the world does not provide for all their needs, Hobbes argued that in the state of nature, without civil society, there will be competition between men for wealth, security, and glory. The ensuing struggle is Hobbes's famous "war of all against all," in which there can be no industry, commerce, or civilization, and the life of man is "solitary, poor, nasty, brutish and short." The struggle occurs because each individual rationally pursues his or her own interests, but the outcome is in no one's interest.

How can this disastrous situation be ended? Not by an appeal to morality or justice; in the state of nature these ideas have no meaning. Yet, we want to survive and we can reason. Our reason leads us to seek peace if it is attainable but to continue to use all the means of war if it is not. How is peace to be obtained? Only by a social contract. We must all agree to give up our rights to attack others in return for their giving up their rights to attack us. By reasoning in order to increase our prospects for survival, we have found the solution.

We know that a social contract will solve our problems. Our reason therefore leads us to desire such an arrangement. But how is it to come about? My reason cannot tell me to accept it while others do not. Nor is Hobbes under the illusion that the mere making of a promise or contract will carry any weight. Since we are self-interested, we will keep our promises only if it is in our interest to do so. A promise that cannot be enforced is worthless. Therefore, in making the social contract, we must establish some means of enforcing it. To do this we must all hand our powers over to some other person or group of persons who will punish anyone who breaches the contract. This person or group of persons Hobbes calls the sovereign. It may be a single person, or an elected legislature, or almost any other form of government; the essence of sovereignty consists only in having sufficient power to keep the peace by punishing those who would break it. When such a sovereign—the Leviathan of his title—exists, justice becomes meaningful in that agreements or promises are necessarily kept. At the same time, each individual has adequate reason to be just, for the sovereign will ensure that those who do not keep their agreements are suitably punished.

Hobbes witnessed the turbulence and near anarchy of the English Civil Wars (1642–51) and was keenly aware of the dangers caused by disputed sovereignty. His solution was to insist that sovereignty must not be divided. Because the sovereign was appointed to enforce the social contract fundamental to peace and everything desired, it can only be rational to resist the sovereign if the sovereign directly threatens one's life. Hobbes was, in effect, a supporter of absolute sovereignty, and this has been the focus of much political discussion of his ideas. His significance for ethics, however, lies rather in his success in dealing with the subject independently of theology and of those quasi-theological or quasi-Aristotelian accounts that see the world as designed for the benefit of human beings. With this achievement, he brought ethics into the modern era.

Early intuitionists: Cudworth, More, and Clarke. There was, of course, immediate opposition to Hobbes's views. Ralph Cudworth (1617–88), one of a group known as the Cambridge Platonists, defended a position in some respects similar to that of Plato. That is to say, Cudworth believed the distinction between good and evil does not lie in human desires but is something objective and can be known by reason, just as the truths of mathematics can be known by reason. Cudworth was thus a forerunner of what has since come to be called intuitionism, the view that there are objective moral truths that can be known by a kind of rational intuition. This view was to attract the support of a line of distinguished thinkers until the 20th century when it became for a time the dominant view in British academic philosophy.

Henry More (1614–87), another leading member of the Cambridge Platonists, attempted to give effect to the comparison between mathematics and morality by listing moral axioms that can be seen as self-evidently true, just as the axioms of geometry are seen to be self-evident. In marked contrast to Hobbes, More included an axiom of benevolence: "If it be good that one man should be supplied with the means of living well and happily, it is mathematically certain that it is doubly good that two should be so supplied, and so on." Here, More was attempting to build on something that Hobbes himself accepted—namely, our own desire to be supplied with the means of living well. More, however, wanted to enlist reason to lead us beyond this narrow egoism to a universal benevolence. There are traces of this line of thought in the Stoics, but it was More who introduced it into British ethical thinking, wherein it is still very much alive.

More's axiom of benevolence

Samuel Clarke (1675–1729), the next major intuitionist, accepted More's axiom of benevolence in slightly different words. He was also responsible for a principle of equity, which, though derived from the Golden Rule so widespread in ancient ethics, was formulated with a new precision: "Whatever I judge reasonable or unreasonable for another to do for me, that by the same judgment I declare reasonable or unreasonable that I in the like case should do for him." As for the means by which these moral truths are known, Clarke accepted Cudworth's and

More's analogy with truths of mathematics and added the idea that what human reason discerns is a certain "fitness or unfitness" about the relationship between circumstances and actions. The right action in a given set of circumstances is the fitting one; the wrong action is unfitting. This is something known intuitively; it is self-evident.

Clarke's notion of fitness is obscure, but intuitionism faces a still more serious problem that has always been a barrier to its acceptance. Suppose we accept the ability of reason to discern that it would be wrong to deceive a person in order to profit from the deception. Why should our discerning this truth provide us with a motive sufficient to override our desire to profit? The intuitionist position divorces our moral knowledge from the forces that motivate us. The former is a matter of reason, the latter of desire.

The punitive power of Hobbes's sovereign is, of course, one way to provide sufficient motivation for obedience to the social contract and to the laws decreed by the sovereign as necessary for the peaceful functioning of society. The intuitionists, however, wanted to show that morality is objective and holds in all circumstances whether there is a sovereign or not. Reward and punishment in the afterlife, administered by an all-powerful God, would provide a more universal motive; and some intuitionists, such as Clarke, did make use of this divine sanction. Other thinkers, however, wanted to show that it is reasonable to do what is good independently of the threats of any external power, human or divine. This desire lay behind the development of the major alternative to intuitionism in 17th- and 18th-century British moral philosophy: moral sense theory. The debate between the intuitionist and moral sense schools of thought aired for the first time the major issue in what is still the central debate in moral philosophy: Is morality based on reason or on feelings?

Shaftesbury and the moral sense school. The term moral sense was first used by the 3rd Earl of Shaftesbury (1671–1713), whose writings reflect the optimistic tone both of the school of thought he founded and of so much of the philosophy of the 18th-century Enlightenment. Shaftesbury believed that Hobbes had erred by presenting a one-sided picture of human nature. Selfishness is not the only natural passion. We also have natural feelings directed to others: benevolence, generosity, sympathy, gratitude, and so on. These feelings give us an "affection for virtue," which leads us to promote the public interest. Shaftesbury called this affection the moral sense, and he thought it created a natural harmony between virtue and self-interest. Shaftesbury was, of course, realistic enough to acknowledge that we also have contrary desires and that not all of us are virtuous all of the time. Virtue could, however, be recommended because—and here Shaftesbury picked up a theme of Greek ethics—the pleasures of virtue are superior to the pleasures of vice.

Harmony between virtue and self-interest through moral sense

Butler on self-interest and conscience. Joseph Butler (1692–1752), a bishop of the Church of England, developed Shaftesbury's position in two ways. He strengthened the case for a harmony between morality and enlightened self-interest by claiming that happiness occurs as a by-product of the satisfaction of desires for things other than happiness itself. Those who aim directly at happiness do not find it; those who have their goals elsewhere are more likely to achieve happiness as well. Butler was not doubting the reasonableness of pursuing one's own happiness as an ultimate aim. He went so far as to say that "... when we sit down in a cool hour, we can neither justify to ourselves this or any other pursuit, till we are convinced that it will be for our happiness, or at least not contrary to it." He held, however, that direct and simple egoism is a self-defeating strategy. Egoists will do better for themselves by adopting immediate goals other than their own interests and living their everyday life in accordance with these more immediate goals.

Butler's second addition to Shaftesbury's account was the idea of conscience. This he saw as a second natural guide to conduct, alongside enlightened self-interest. Butler believed that there is no inconsistency between the two; he admitted, however, that skeptics may doubt "the happy tendency of virtue" and for them conscience can serve as an authoritative guide. Just what reason these skeptics have to follow conscience, if they believe its guidance to be contrary to their own happiness, is something that Butler did not adequately explain. Nevertheless, his introduction of conscience as an independent source of moral reasoning reflects an important difference between ancient and modern ethical thinking. The Greek and Roman philosophers would have had no difficulty in accepting everything Butler said about the pursuit of happiness, but they would not have understood his idea of another independent source of rational guidance. Although Butler insisted that the two operate in harmony, this was for him a fortunate fact about the world and not a necessary principle of reason. Thus his recognition of conscience opened the way for later formulations of a universal principle of conduct at odds with the path indicated by even the most enlightened self-interested reasoning.

Conscience is a rational guide to conduct

The climax of moral sense theory: Hutcheson and Hume. The moral sense school reached its fullest development in the works of two Scottish philosophers, Francis Hutcheson (1694–1746) and David Hume (1711–76). Hutcheson was concerned with showing, against the intuitionists, that moral judgment cannot be based on reason and therefore must be a matter of whether an action is "amiable or disagreeable" to one's moral sense. Like Butler's notion of conscience, Hutcheson's moral sense does not find pleasing only, or even predominantly, those actions that are in one's own interest. On the contrary, Hutcheson conceived moral sense as based on a disinterested benevolence. This led him to state, as the ultimate criterion of the goodness of an action, a principle that was to serve as the basis for the Utilitarian reformers: "that action is best which procures the greatest happiness for the greatest numbers . . . ."

Hume, like Hutcheson, held that reason cannot be the basis of morality. His chief ground for this conclusion was that morality is essentially practical: there is no point in judging something good if the judgment does not incline us to act accordingly. Reason alone, however, Hume regarded as "the slave of the passions." Reason can show us how best to achieve our ends, but it cannot determine our ultimate desires and is incapable of moving us to action except in accordance with some prior want or desire. Hence, reason cannot give rise to moral judgments.

This is an important argument that is still employed in the debate between those who believe that morality is based on reason and those who base it instead on emotion or feelings. Hume's conclusion certainly follows from his premises. Can either premise be denied? We have seen that intuitionists such as Cudworth and Clarke maintained that reason can lead to action. Reason, they would have said, leads us to see a particular action as fitting in given circumstances and therefore to do it. Hume would have none of this. "Tis not contrary to reason," he provocatively asserted, "to prefer the destruction of the whole world to the scratching of my finger." To show that he was not embracing the view that only egoism is rational, Hume continued: "Tis not contrary to reason to choose my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me." His point was simply that to have these preferences is to have certain desires or feelings; they are not matters of reason at all. The intuitionists might insist that moral and mathematical reasoning are analogous, but this analogy was not helpful here. We can know a truth of geometry and not be motivated to act in any way.

What of Hume's other premise that morality is essentially practical and moral judgments must lead to action? This can be denied more easily. We could say that moral judgments merely tell us what is right or wrong. They do not lead to action unless we want to do what is right. Then Hume's argument would do nothing to undermine the claim that moral judgments are based on reason. But there is a price to pay. The terms right and wrong lose much of their force. We can no longer assert that those who know what is right but do what is wrong are in any way irrational. They are just people who do not happen to have the desire to do what is right. This desire—because it leads to action—must be acknowledged to be based on feeling rather than reason. Denying that morality is necessarily action-guiding means abandoning the idea, so important

to those defending the objectivity of morality, that some things are objectively required of all rational beings.

Hume's forceful presentation of this argument against a rational basis for morality would have been enough to earn him a place in the history of ethics, but it is by no means his only achievement in this field. In *A Treatise of Human Nature* (1739–40) Hume points, almost as an afterthought, to the fact that writers on morality regularly start by making various observations about human nature or about the existence of a god—all statements of fact about what is the case—and then suddenly switch to statements about what ought or ought not be done. Hume says that he cannot conceive how this new relationship of "ought" can be deduced from the preceding statements that were related by "is"; and he suggests these authors should explain how this deduction is to be achieved. The point has since been called Hume's Law and taken as proof of the existence of a gulf between facts and values, or between "is" and "ought." This places too much weight on Hume's brief and ironic comment, but there is no doubt that many writers, both before and after Hume, have argued as if values could easily be deduced from facts. They can usually be found to have smuggled values in somewhere. Attention to Hume's Law makes it easy for us to detect such logically illicit contraband.

Hume's Law

Hume's positive account of morality is in line with that of the moral sense school: "The hypothesis which we embrace is plain. It maintains that morality is determined by sentiment. It defines virtue to be whatever mental action or quality gives to a spectator the pleasing sentiment of approbation; and vice the contrary." In other words, Hume takes moral judgments to be based on a feeling. They do not reflect any objective state of the world. Having said that, however, it may still be asked whether this feeling is one that is common to all of us or one that varies from individual to individual. If Hume gives the former answer, moral judgments retain a kind of objectivity. While they do not reflect anything out there in the universe apart from human feelings, one's judgments may be true or false depending on whether they capture this universal human moral sentiment. If, on the other hand, the feeling varies from one individual to the next, moral judgments become entirely subjective. People's judgments would express their own feelings, and to reject someone else's judgment as wrong would merely be to say that one's own feelings were different.

Hume does not make entirely clear which of these two views he holds; but if he is to avoid breaching his own rule about not deducing an "ought" from an "is," he cannot hold that a moral judgment can follow logically from a description of the feelings that an action gives to a particular group of spectators. From the mere existence of a feeling we cannot draw the inference that we ought to obey it. For Hume to be consistent on this point—and even with his central argument that moral judgments must move to action—the moral judgment must be based not on the fact that all people, or most people, or even the speaker, have a certain feeling; it must rather be based on the actual experience of the feeling by whoever accepts the judgment. This still leaves it open whether the feeling is common to all or limited to the person accepting the judgment, but it shows that, in either case, the "truth" of a judgment for any individual depends on whether that individual actually has the appropriate feeling. Is this "truth" at all? As will be seen below, 20th-century philosophers with views broadly similar to Hume's have suggested that moral judgments have a special kind of meaning not susceptible of truth or falsity in the ordinary way.

The intuitionist response: Price and Reid. Powerful as they were, Hume's arguments did not end the debate between the moral sense theorists and the intuitionists. They did, however, lead Richard Price (1723–91), Thomas Reid (1710–96), and later intuitionists to abandon the idea that moral truths can be established by some process of demonstrative reasoning akin to that used in mathematics. Instead, these proponents of intuitionism took the line that our notions of right and wrong are simple, objective ideas, directly perceived by us and not further analyzable into anything such as "fitness." We know of these ideas, not through any moral sense based on feelings, but rather through a faculty of reason or of the intellect that is capable of discerning truth. Since Hume, this has been the only plausible form of intuitionism. Yet, Price and Reid failed to explain adequately just what are the objective moral qualities that we perceive directly and how they connect with the actions we choose.

Utilitarianism. At this point the argument over whether morality is based on reason or feelings was temporarily exhausted, and the focus of British ethics shifted from such questions about the nature of morality as a whole to an inquiry into which actions are right and which are wrong. Today, the distinction between these two types of inquiry would be expressed by saying that whereas the 18th-century debate between intuitionism and the moral sense school dealt with questions of metaethics, 19th-century thinkers became chiefly concerned with questions of normative ethics. The positions we take in metaethics over whether ethics is objective or subjective, for example, do not tell us what we ought to do. That task is the province of normative ethics.

Paley. The impetus to the discussion of normative ethics was provided by the challenge of Utilitarianism. The essential principle of Utilitarianism was, as noted above, put forth by Hutcheson. Curiously, it gained further development from the widely read theologian William Paley (1743–1805), who provides a good example of the independence of metaethics and normative ethics. His position on the nature of morality was similar to that of Ockham and Luther—namely, he held that right and wrong are determined by the will of God. Yet, because he believed that God wills the happiness of his creatures, his normative ethics were Utilitarian: whatever increases happiness is right; whatever diminishes it is wrong.

Bentham. Notwithstanding these predecessors, Jeremy Bentham (1748–1832) is properly considered the father of modern Utilitarianism. It was he who made the Utilitarian principle serve as the basis for a unified and comprehensive ethical system that applies, in theory at least, to every area of life. Never before had a complete, detailed system of ethics been so consistently constructed from a single fundamental ethical principle.

Bentham as the father of modern Utilitarianism

Bentham's ethics began with the proposition that nature has placed human beings under two masters: pleasure and pain. Anything that seems good must either be directly pleasurable, or thought to be a means to pleasure or to the avoidance of pain. Conversely, anything that seems bad must either be directly painful, or thought to be a means to pain or to the deprivation of pleasure. From this Bentham argued that the words right and wrong can only be meaningful if they are used in accordance with the Utilitarian principle, so that whatever increases the net surplus of pleasure over pain is right and whatever decreases it is wrong.

Bentham then set out how we are to weigh the consequences of an action, and thereby decide whether it is right or wrong. We must, he says, take account of the pleasures and pains of everyone affected by the action, and this is to be done on an equal basis: "Each to count for one, and none for more than one." (At a time when Britain had a major trade in slaves, this was a radical suggestion; and Bentham went further still, explicitly extending consideration to nonhuman animals as well.) We must also consider how certain or uncertain the pleasures and pains are, their intensity, how long they last, and whether they tend to give rise to further feelings of the same or of the opposite kind.

Bentham did not allow for distinctions in the quality of pleasure or pain as such. Referring to a popular game, he affirmed that "quantity of pleasure being equal, pushpin is as good as poetry." This led his opponents to characterize his philosophy as one fit for pigs. The charge is only half true. Bentham could have defended a taste for poetry on the grounds that whereas one tires of mere games, the pleasures of a true appreciation of poetry have no limit; thus the quantities of pleasure obtained by poetry are greater than those obtained by pushpin. All the same, one of the strengths of Bentham's position is its honest bluntness, which it owes to his refusal to be fazed by the

contrary opinions either of conventional morality or of refined society. He never thought that the aim of Utilitarianism was to explain or justify ordinary moral views; it was, rather, to reform them.

Mill. John Stuart Mill (1806–73), Bentham's successor as the leader of the Utilitarians and the most influential British thinker of the 19th century, had some sympathy for the view that Bentham's position was too narrow and crude. His essay "Utilitarianism" (1861) introduced several modifications, all aimed at a broader view of what is worthwhile in human existence and at implications less shocking to established moral convictions. Although his position was based on the maximization of happiness (and this is said to consist in pleasure and the absence of pain), he distinguished between pleasures that are higher and those that are lower in quality. This enabled him to say that it is "better to be Socrates dissatisfied than a fool satisfied." The fool, he argued, would only be of a different opinion because he did not know both sides of the question.

Mill sought to show that Utilitarianism is compatible with moral rules and principles relating to justice, honesty, and truthfulness by arguing that Utilitarians should not attempt to calculate before each action whether that specific action will maximize utility. Instead, they should be guided by the fact that an action falls under a general principle (such as the principle that we should keep our promises), and adherence to that general principle tends to increase happiness. Only under special circumstances is it necessary to consider whether an exception may have to be made.

Sidgwick. Mill's easily readable prose ensured a wide audience for his exposition of Utilitarianism, but as a philosopher he was markedly inferior to the last of the 19th-century Utilitarians, Henry Sidgwick (1838–1900). Sidgwick's *Methods of Ethics* (1874) is the most detailed and subtle work of Utilitarian ethics yet produced. Especially noteworthy is his discussion of the various principles accepted by what he calls common sense morality—*i.e.,* the morality accepted by most people without systematic thought. Price, Reid, and some adherents of their brand of intuitionism thought that such principles (*e.g.,* those of truthfulness, justice, honesty, benevolence, purity, and gratitude) were self-evident, independent moral truths. Sidgwick was himself an intuitionist as far as the basis of ethics was concerned: he believed that the principle of Utilitarianism must ultimately be based on a self-evident axiom of rational benevolence. Nonetheless, he strongly rejected the view that all principles of common sense morality are themselves self-evident. He went on to demonstrate that the allegedly self-evident principles conflict with one another and are vague in their application. They could only be part of a coherent system of morality, he argued, if they were regarded as subordinate to the Utilitarian principle, which defined their application and resolved the conflicts between them.

Attempt at reconciling common sense morality and Utilitarianism

Sidgwick was satisfied that he had reconciled common sense morality and Utilitarianism by showing that whatever was sound in the former could be accounted for by the latter. He was, however, troubled by his inability to achieve any such reconciliation between Utilitarianism and egoism, the third method of ethical reasoning dealt with in his book. True, Sidgwick regarded it as self-evident that "from the point of view of the universe" one's own good is of no greater value than the like good of any other person, but what could be said to the egoist who expresses no concern for the point of view of the universe, taking his stand instead on the fact that his own good mattered more to him than anyone else's? Bentham had apparently believed either that self-interest and the general happiness are not at odds or that it is the legislator's task to reward or punish actions so as to see that they are not. Mill also had written of the need for sanctions but was more concerned with the role of education in shaping human nature in such a way that one finds happiness in doing what benefits all. By contrast, Sidgwick was convinced that this could lead at best to a partial overlap between what is in one's own interest and what is in the interest of all. Hence, he searched for arguments with which to convince the egoist of the rationality of universal benevolence but failed to find any. *The Methods of Ethics* concludes with an honest admission of this failure and an expression of dismay at the fact that, as a result, ". . . it would seem necessary to abandon the idea of rationalizing [morality] completely."

THE CONTINENTAL TRADITION: FROM SPINOZA TO NIETZSCHE

Spinoza. If Hobbes is to be regarded as the first of a distinctively British philosophical tradition, the Dutch-Jewish philosopher Benedict Spinoza (1632–77) appropriately occupies the same position in continental Europe. Unlike Hobbes, Spinoza did not provoke a long-running philosophical debate. In fact, his philosophy was neglected for a century after his death and was in any case too much of a self-contained system to invite debate. Nevertheless, Spinoza held positions on crucial issues that were in sharp contrast to those taken by Hobbes, and these differences were to grow over the centuries during which British and continental European philosophy followed their own paths.

The first of these contrasts with Hobbes is Spinoza's attitude toward natural desires. As has been noted, Hobbes took self-interested desire for pleasure as an unchangeable fact about human nature and proceeded to build a moral and political system to cope with it. Spinoza did just the opposite. He saw natural desires as a form of bondage. We do not choose to have them of our own will. Our will cannot be free if it is subject to forces outside itself. Thus our real interests lie not in satisfying these desires but in transforming them by the application of reason. Spinoza thus stands in opposition not only to Hobbes but also to the position later to be taken by Hume, for Spinoza saw reason not as the slave of the passions but as their master.

The second important contrast is that while individual humans and their separate interests are always assumed in Hobbes's philosophy, this separation is simply an illusion from Spinoza's viewpoint. Everything that exists is part of a single system, which is at the same time nature and God. (One possible interpretation of this is that Spinoza was a pantheist, believing that God exists in every aspect of the world and not apart from it.) We, too, are part of this system and are subject to its rationally necessary laws. Once we know this, we understand how irrational it would be to desire that things should be different from the way they are. This means that it is irrational to envy, to hate, and to feel guilt, for these emotions presuppose the possibility of things being different. So we cease to feel such emotions and find peace, happiness, and even freedom—in Spinoza's terms the only freedom there can be—in understanding the system of which we are a part.

A view of the world so different from our everyday conceptions as that of Spinoza's cannot be made to seem remotely plausible when presented in summary form. To many philosophers it remains implausible even when complete. Its value for ethics, however, lies not in its validity as a whole, but in the introduction into continental European philosophy of a few key ideas: that our everyday nature may not be our true nature; that we are part of a larger unity; and that freedom is to be found in following reason.

Leibniz. The German philosopher and mathematician Gottfried Wilhelm Leibniz (1646–1716), the next great figure in the Rationalist tradition, gave scant attention to ethics, perhaps because of his belief that the world is governed by a perfect God, and hence must be the best of all possible worlds. As a result of Voltaire's hilarious parody in *Candide* (1758), this position has achieved a certain notoriety. It is not generally recognized, however, that it does at least provide a consistent solution to a problem that has baffled thinking Christians for many centuries: How can there be evil in a world governed by an all-powerful, all-knowing, and all-good God? Leibniz's solution may not be plausible, but there may be no better one if the above premises are allowed to pass unchallenged.

Rousseau. It was the French philosopher and writer Jean-Jacques Rousseau (1712–78) who took the next step. His *Discours sur l'origine et les fondements de l'inégalité parmi les hommes* (1755; *A Discourse upon the Origin and*

Foundation of the Inequality Among Mankind) depicted a state of nature very different from that described by Hobbes as well as from Christian conceptions of original sin. Rousseau's "noble savages" lived isolated, trouble-free lives, supplying their simple wants from the abundance that nature provided and even coming to each other's aid in times of need. Only when someone claimed possession of a piece of land did laws have to be introduced, and with them came civilization and all its corrupting influences. This is, of course, a message that resembles one of Spinoza's key points: The human nature we see before us in our fellow citizens is not the only possibility; somewhere, there is something better. If we can find a way to reach it, we will have found the solution to our ethical and social problems.

Rousseau revealed his route in his *Contrat social* (1762; *A Treatise on the Social Compact*, or *Social Contract*). It required rule by the "general will." This may sound like democracy and, in a sense, it was democracy that Rousseau advocated; but his conception of rule by the general will is very different from the modern idea of democratic government. Today, we assume that in any society the interests of different citizens will be in conflict, and that as a result for every majority that succeeds in having its will implemented there will be a minority that fails to do so. For Rousseau, on the other hand, the general will is not the sum of all the individual wills in the community but the true common will of all the citizens. Even if a person dislikes and opposes a decision carried by the majority, that decision represents the general will, the common will in which he shares. For this to be possible, Rousseau must be assuming that there is some common good in which all human beings share and hence that their true interests coincide. As man passes from the state of nature to civil society, he has to "consult his reason rather than study his inclinations." This is not, however, a sacrifice of his true interests, for in following reason he ceases to be a slave to "physical impulses" and so gains moral freedom.

Rousseau's notion of general will

This leads to a picture of civilized human beings as divided selves. The general will represents the rational will of every member of the community. If an individual opposes the decision of the general will, his opposition must stem from his physical impulses and not from his true, autonomous will. For obvious reasons, this idea was to find favour with such autocratic leaders of the French Revolution as Robespierre. It also had a much less sinister influence on one of the outstanding philosophers of modern times: Immanuel Kant of Germany.

Kant. Interestingly, Kant (1724–1804) acknowledged that he had despised the ignorant masses until he read Rousseau and came to appreciate the worth that exists in every human being. For other reasons too, Kant is part of the tradition deriving from both Spinoza and Rousseau. Like his predecessors, Kant insisted that actions resulting from desires cannot be free. Freedom is to be found only in rational action. Moreover, whatever is demanded by reason must be demanded of all rational beings; hence, rational action cannot be based on a single individual's personal desires, but must be action in accordance with something that he can will to be a universal law. This view roughly parallels Rousseau's idea of the general will as that which, as opposed to the individual will, a person shares with the whole community. Kant extended this community to all rational beings.

Kant's most distinctive contribution to ethics was his insistence that our actions possess moral worth only when we do our duty for its own sake. He first introduced this idea as something accepted by our common moral consciousness and only then tried to show that it is an essential element of any rational morality. In claiming that this idea is central to the common moral consciousness, Kant was expressing in heightened form a tendency of Judeo-Christian ethics and revealing how much the Western ethical consciousness had changed since the time of Socrates, Plato, and Aristotle.

Does our common moral consciousness really insist that there is no moral worth in any action done for any motive other than duty? Certainly we would be less inclined to praise the young man who plunges into the surf to rescue a drowning child if we learned that he did it because he expected a handsome reward from the child's millionaire father. This feeling lies behind Kant's disagreement with all those moral philosophers who have argued that we should do what is right because that is the path to happiness, either on earth or in heaven. But Kant went further than this. He was equally opposed to those who see benevolent or sympathetic feelings as the basis of morality. Here he may be reflecting the moral consciousness of 18th-century Protestant Germany, but it appears that even then the moral consciousness of Britain, as reflected in the writings of Shaftesbury, Hutcheson, Butler, and Hume, was very different. The moral consciousness of Western civilization in the last quarter of the 20th century also appears to be different from the one Kant was describing.

Kant's ethics is based on his distinction between hypothetical and categorical imperatives. He called any action based on desires a hypothetical imperative, meaning by this that it is a command of reason that applies only if we desire the goal. For example, "Be honest, so that people will think well of you!" is an imperative that applies only if you want people to think well of you. A similarly hypothetical analysis can be given of the imperatives suggested by, say, Shaftesbury's ethics: "Help those in distress, if you sympathize with their sufferings!" In contrast to such approaches to ethics, Kant said that the commands of morality must be categorical imperatives: they must apply to all rational beings, regardless of their wants and feelings. To most philosophers this poses an insuperable problem: a moral law that applied to all rational beings, irrespective of their personal wants and desires, could have no specific goals or aims because all such aims would have to be based on someone's wants or desires. It took Kant's peculiar genius to seize upon precisely this implication, which to others would have refuted his claims, and to use it to derive the nature of the moral law. Because nothing else but reason is left to determine the content of the moral law, the only form this law can take is the universal principle of reason. Thus the supreme formal principle of Kant's ethics is: "Act only on that maxim through which you can at the same time will that it should become a universal law."

The dictates of morality as categorical imperatives

Kant still faced two major problems. First, he had to explain how we can be moved by reason alone to act in accordance with this supreme moral law; and, second, he had to show that this principle is able to provide practical guidance in our choices. If we were to couple Hume's theory that reason is always the slave of the passions with Kant's denial of moral worth to all actions motivated by desires, the outcome would be that no actions can have moral worth. To avoid such moral skepticism, Kant maintained that reason alone can lead to action. Unfortunately he was unable to say much in defense of this claim. Of course, the mere fact that we otherwise face so unpalatable a conclusion is in itself a powerful incentive to believe that somehow a categorical imperative must be possible, but this is not convincing to anyone not already wedded to Kant's view of moral worth. At one point Kant appeared to be taking a different line. He wrote that the moral law inevitably produces in us a feeling of reverence or awe. If he meant to say that this feeling then becomes the motivation for obedience, however, he was conceding Hume's point that reason alone is powerless to bring about action. It would also be difficult to accept that anything, even the moral law, can necessarily produce a certain kind of feeling in all rational beings regardless of their psychological constitution. Thus this approach does not succeed in clarifying Kant's position or rendering it plausible.

Kant gave closer attention to the problem of how his supreme formal principle of morality can provide guidance in concrete situations. One of his examples is as follows. Suppose that I plan to get some money by promising to pay it back, although I have no intention of keeping my promise. The maxim of such an action might be "Make false promises when it suits you to do so." Could such a maxim be a universal law? Of course not. If promises were so easily broken, no one would rely on them, and the practice of promising would cease. For this reason,

I know that the moral law does not allow me to carry out my plan.

Not all situations are so easily decided. Another of Kant's examples deals with aiding those in distress. I see someone in distress, whom I could easily help, but I prefer not to do so. Can I will as a universal law the maxim that a person should refuse assistance to those in distress? Unlike the case of promising, there is no strict inconsistency in this maxim being a universal law. Kant, however, says that I cannot will it to be such because I may someday be in distress myself, and I would then want assistance from others. This type of example is less convincing than the previous one. If I value self-sufficiency so highly that I would rather remain in distress than escape from it through the intervention of another, Kant's principle no longer tells me that I have a duty to assist those in distress. In effect, Kant's supreme principle of practical reason can only tell us what to do in those special cases in which turning the maxim of our action into a universal law yields a contradiction. Outside this limited range, the moral law that was to apply to all rational beings regardless of their wants and desires cannot guide us except by appealing to our desires.

Kant does offer alternative formulations of the categorical imperative, and one of these has been seen as providing more substantial guidance than the formulation so far considered. This formulation is: "So act that you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as means." The connection between this formulation and the first one is not entirely clear, but the idea seems to be that when I choose for myself I treat myself as an end. If, therefore, in accordance with the principle of universal law, I must choose so that all could choose similarly, I must respect everyone else as an end. Even if this is valid, the application of the principle raises further questions. What is it to treat someone merely as a means? Using a person as a slave is an obvious example; Kant, like Bentham, was making a stand against this kind of inequality while it still flourished as an institution in some parts of the world. But to condemn slavery we have only to give equal weight to the interests of the slaves. Does Kant's principle take us any further than Utilitarianism? Modern Kantians hold that it does because they interpret it as denying the legitimacy of sacrificing the rights of one human being in order to benefit others.

The deontological nature of Kant's ethics

One thing that can be said confidently is that Kant was firmly opposed to the Utilitarian principle of judging every action by its consequences. His ethics is a deontology. In other words, the rightness of an action depends on whether it accords with a rule irrespective of its consequences. In one essay Kant went so far as to say that it would be wrong to tell a lie even to a would-be murderer who came to your door seeking to kill an innocent person hidden in your house. This kind of situation illustrates how difficult it is to remain a strict deontologist when principles may clash. Apparently Kant believed that his principle of universal law required that one never tell lies, but it could also be argued that his principle of treating everyone as an end would necessitate doing everything possible to save the life of an innocent person. Another possibility would be to formulate the maxim of the action with sufficient precision to define the circumstances under which it would be permissible to tell lies—*e.g.*, we could all agree to a universal law that permitted lies to people intending to commit murder. Kant did not explore such solutions.

Hegel. Kant's philosophy deeply affected subsequent German thought, but there were several aspects of it that troubled later thinkers. One of these was his portrayal of human nature as irreconcilably split between reason and emotion. In *Briefe über die ästhetische Erziehung des Menschen* (1795; *Letters on the Aesthetic Education of Man*), the dramatist and literary theorist Friedrich Schiller suggested that while this might apply to modern human beings, it was not the case in ancient Greece where reason and feeling seemed to have been in harmony. (There is, as suggested earlier, some basis for this claim insofar as the Greek moral consciousness did not make the modern distinction between morality and self-interest.) Schiller's suggestion may have been the spark that led Georg Wilhelm Friedrich Hegel (1770–1831) to develop the first philosophical system that has historical change as its core.

As Hegel presents it, all of history is the progress of mind or spirit along a logically necessary path that leads to freedom. Human beings are manifestations of this universal mind, although at first they do not realize this. Freedom cannot be achieved until human beings do realize it, and so feel at home in the universe. There are echoes of Spinoza in Hegel's idea of mind as something universal and also in his conception of freedom as based on knowledge. What is original, however, is the way in which all of history is presented as leading to the goal of freedom. Thus Hegel accepts Schiller's view that for the ancient Greeks, reason and feeling were in harmony, but he sees this as a naive harmony that could exist only as long as the Greeks did not see themselves as free individuals with a conscience independent of the views of the community. For freedom to develop, it was necessary for this harmony to break down. This occurred as a result of the Reformation, with its insistence on the right of individual conscience. But the rise of individual conscience left human beings divided between conscience and self-interest, between reason and feeling. We have seen how many philosophers tried unsuccessfully to bridge this gulf until Kant's insistence that we must do our duty for duty's sake made the division an apparently inevitable part of moral life. For Hegel, however, it can be overcome by a synthesis of the harmonious communal nature of Greek life with the modern freedom of individual conscience.

The concept of the organic community

In *Naturrecht und Staatswissenschaft im Grundrisse*, alternatively entitled *Grundlinien der Philosophie des Rechts* (1821; *The Philosophy of Right*), Hegel described how this synthesis could be achieved in an organic community. The key to his solution is the recognition that human nature is not fixed but is shaped by the society in which one lives. The organic community would foster those desires that most benefit the community. It would imbue its members with the sense that their own identity consists in being a part of the community, so that they would no more think of going off in pursuit of their own private interests than one's left arm would think of going off without the rest of the body. Nor should it be forgotten that such organic relationships are reciprocal: the organic community will no more disregard the interests of its members than an individual would disregard an injury to his or her arm. Harmony would thus prevail but not the naive harmony of ancient Greece. The citizens of Hegel's organic community do not obey its laws and customs simply because they are there. With the independence of mind characteristic of modern times, they can only give their allegiance to institutions that they recognize as conforming to rational principles. The modern organic state, unlike the ancient Greek city-state, is self-consciously based on rationally selected principles.

Hegel provided a new approach to the ancient problem of reconciling morality and self-interest. Others had accepted the problem as part of the inevitable nature of things and looked for ways around it. Hegel looked at it historically and saw it as a problem only in a certain type of society. Instead of solving the problem as it existed, he looked to the emergence of a new form of society in which it would disappear. In this way Hegel claimed to have overcome one great problem that was insoluble for Kant.

Hegel also believed that he had the solution to the other key weakness in Kant's ethics—namely, the difficulty of giving content to the supreme formal moral principle. In Hegel's organic community, the content of our moral duty would be given to us by our position in society. We would know that our duty was to be a good parent, a good citizen, a good teacher, merchant, or soldier, as the case might be. It is an ethic that has been called "my station and its duties." It might be thought that this is a limited, conservative conception of what we ought to do with our lives, especially when compared with Kant's principle of universal law, which does not base what we ought to do on what our particular station in society happens to be. Hegel would have replied that because the organic community is

based on universally valid principles of reason, it complies with Kant's principle of universal law. Moreover, without the specific content provided by the concrete institutions and practices of a society, that principle would remain an empty formula.

Hegel's philosophy has both a conservative and a radical side. The conservative aspect is reflected in the ethic of "my station and its duties," and even more strongly in the significant resemblance between Hegel's detailed description of the organic society and the actual institutions of the Prussian state in which he lived and taught for the last decade of his life. This resemblance, however, was in no way a necessary implication of Hegel's philosophy as a whole. After Hegel's death, a group of his more radical followers known as the Young Hegelians hailed the manner in which he had demonstrated the need for a new form of society to overcome the separation between self and community but scorned the implication that the state in which they were living could be this solution to all the problems of history. Among this group was a young student named Karl Marx.

Marx. Marx (1818–83) has often been presented by his followers as a scientist rather than a moralist. He did not deal directly with the ethical issues that occupied the philosophers so far discussed. His Materialist conception of history is, rather, an attempt to explain all ideas, whether political, religious, or ethical, as the product of the particular economic stage that society has reached. Thus a feudal society will regard loyalty and obedience to one's lord as the chief virtues. A capitalist economy, on the other hand, requires a mobile labour force and expanding markets, so that freedom, especially the freedom to sell one's labour, is its key ethical conception. Because Marx saw ethics as a mere by-product of the economic basis of society, he frequently took a dismissive stance toward it. Echoing the Sophist Thrasymachus, Marx said that the "ideas of the ruling class are in every epoch the ruling ideas." With his coauthor Friedrich Engels, he was even more scornful in the *Manifest der Kommunistischen Partei* (1848; *The Communist Manifesto*), in which morality, law, and religion are referred to as "so many bourgeois prejudices behind which lurk in ambush just as many bourgeois interests."

A sweeping rejection of ethics, however, is difficult to reconcile with the highly moralistic tone of Marx's condemnation of the miseries the capitalist system inflicts upon the working class and with his obvious commitment to hastening the arrival of the Communist society that will end such iniquities. After Marx died, Engels tried to explain this apparent inconsistency by saying that as long as society was divided into classes, morality would serve the interests of the ruling class. A classless society, on the other hand, would be based on a truly human morality that served the interests of all human beings. This does make Marx's position consistent by setting him up as a critic, not of ethics as such, but rather of the class-based moralities that would prevail until the Communist revolution.

By studying Marx's earlier writings—those produced when he was a Young Hegelian—one obtains a slightly different, though not incompatible, impression of the place of ethics in Marx's thought. There seems no doubt that the young Marx, like Hegel, saw human freedom as the ultimate goal. He also held, as did Hegel, that freedom could only be obtained in a society in which the dichotomy between private interest and the general interest had disappeared. Under the influence of socialist ideas, however, he formed the view that merely knowing what was wrong with the world would not achieve anything. Only the abolition of private property could lead to the transformation of human nature and so bring about the reconciliation of the individual and the community. Theory, Marx concluded, had gone as far as it could; even the theoretical problems of ethics, as illustrated in Kant's division between reason and feeling, would remain insoluble unless one moved from theory to practice. This is what Marx meant in the famous thesis that is engraved on his tombstone: "The philosophers have only interpreted the world, in various ways; the point is to change it." The goal of changing the world stemmed from Marx's attempt to overcome one of the central problems of ethics; the means now passed beyond philosophy.

Nietzsche. Friedrich Nietzsche (1844–1900) was a literary and social critic, not a systematic philosopher. In ethics, the chief target of his criticism is the Judeo-Christian tradition. He describes Jewish ethics as a "slave morality" based on envy. Christian ethics is, in his opinion, even worse because it makes a virtue of meekness, poverty, and humility, telling one to turn the other cheek rather than to struggle. It is the ethics of the weak, who hate and fear strength, pride, and self-affirmation. Such an ethics undermines the human drives that have led to the greatest and most noble human achievements.

Nietzsche thought the era of traditional religion to be over: "God is dead," perhaps his most widely repeated aphorism, was his paradoxical way of putting it. Yet, what was to be put in its place? Nietzsche took from Aristotle the concept of greatness of soul, the unchristian virtue that included nobility and a justified pride in one's achievements. He suggested a reevaluation of values that would lead to a new ideal: the *Übermensch,* a term usually translated as "Superman" and given connotations that suggest that Nietzsche would have regarded Hitler as an ideal type. Nietzsche's praise of "the will to power" is taken as further evidence that he would have approved of Hitler. This interpretation owes much to Nietzsche's racist sister, who after his death compiled a volume of his unpublished writings, arranging them to make it appear that he was a forerunner of Nazi thinking. This is at best a partial truth. Nietzsche was almost as contemptuous of pan-German racism and anti-Semitism as he was of the ethics of Judaism and Christianity. What Nietzsche meant by *Übermensch* was a person who could rise above the limitations of ordinary morality; and by "the will to power" it seems that Nietzsche had in mind self-affirmation and not necessarily the use of power to oppress others.

The concept of the *Übermensch* and its implications

Nevertheless, Nietzsche left himself wide open to those who wanted his philosophical imprimatur for their crimes against humanity. His belief in the importance of the *Übermensch* made him talk of ordinary people as "the herd," who did not really matter. In *Jenseits von Gut und Böse* (1886; *Beyond Good and Evil*), he wrote with approval of "the distinguished type of morality," according to which "one has duties only toward one's equals; toward beings of a lower rank, toward everything foreign to one, one may act as one sees fit, 'as one's heart dictates' "—in any event, beyond good and evil. The point is that the *Übermensch* is above all ordinary moral standards: "The distinguished type of human being feels *himself* as value-determining; he does not need to be ratified; he judges 'that which is harmful to me is harmful as such'; he knows that *he* is the something which gives value to objects; he *creates values.*" In this Nietzsche was a forerunner of Existentialism rather than Nazism, but then Existentialism, precisely because it gives no basis for choosing other than authenticity, is not incompatible with Nazism.

Nietzsche's position on ethical matters represents a stark contrast to that of Henry Sidgwick, the last major figure of 19th-century British ethics treated in this article. Sidgwick believed in objective standards for ethical judgments and thought that the subject of ethics had over the centuries made progress toward these standards. He saw his own work as building carefully on that progress. Nietzsche, on the other hand, would have us sweep away everything since Greek ethics and not keep much of that either. The superior types would then be able to freely create their own values as they saw fit.

20th-century Western ethics

The brief historical survey of Western ethics from Socrates to the 20th century provided above has shown three constant themes. Since the Sophists, there have been (1) disagreements over whether ethical judgments are truths about the world or only reflections of the wishes of those who make them; (2) frequent attempts to show, in the face of considerable skepticism, either that it is in one's own interests to do what is good or that, even though this is not necessarily in one's own interests, it is the ratio-

nal thing to do; and (3) repeated debates over just what goodness and the standard of right and wrong might be. The 20th century has seen new twists to these old themes and an increased attention to the application of ethics to practical problems. Each of these major questions is considered below in terms of metaethics, normative ethics, and applied ethics.

METAETHICS

As previously noted, metaethics deals not with substantive ethical theories or moral judgments but rather with questions about the nature of these theories and judgments. Among 20th-century philosophers in English-speaking countries, those defending the objectivity of ethical judgments have most often been intuitionists or naturalists; those taking a different view have been emotivists or prescriptivists.

Moore and the naturalistic fallacy. At first it was the intuitionists who dominated the scene. In 1903 the Cambridge philosopher G.E. Moore presented in *Principia Ethica* his "open question argument" against what he called the naturalistic fallacy. The argument can in fact be found in Sidgwick and to some extent in the 18th-century intuitionists, but Moore's statement of it somehow caught the imagination of philosophers for the first half of the 1900s. Moore's aim was to prove that "good" is the name of a simple, unanalyzable quality. His chief target was the attempt to define good in terms of some natural quality of the world whether it be "pleasure" (he had John Stuart Mill in mind), or "more evolved" (here he refers to Herbert Spencer, who had tried to build an ethical system around Darwin's theory of evolution), or simply the idea of what is natural itself, as in appeals to a law of nature—hence the label naturalistic fallacy (*i.e.*, the fallacy of treating good as if it were the name of a natural property). But the label is not apt because Moore's argument applied, as he acknowledged, to any attempt to define good in terms of something else, including something metaphysical or supernatural such as "what God wills."

The naturalistic fallacy

The so-called open question argument itself is simple enough. It consists of taking the proposed definition of good and turning it into a question. For instance, if the proposed definition is "Good means whatever leads to the greatest happiness of the greatest number," then Moore would ask: "Is whatever leads to the greatest happiness of the greatest number good?" Moore is not concerned whether we answer yes or no. His point is that if the question is at all meaningful—if a negative answer is not plainly self-contradictory—then the definition cannot be right, for a definition is supposed to preserve the meaning of the term defined. If it does, a question of the type Moore asks would be absurd for all who understand the meaning of the term. Compare, for example, "Do all squares have four equal sides?"

Moore's argument does show that definitions of the kind he criticized do not capture all that we ordinarily mean by the term good. It would still be open to a would-be naturalist to admit that the definition does not capture everything that we ordinarily mean by the term, and add that all this shows is that ordinary usage is muddled and in need of revision. (We shall see that J.L. Mackie was later to make this part of his defense of subjectivism.) As for Mill, it is questionable whether he really intended to offer a definition of the term good; he seems to have been more interested in offering a criterion by which we could ascertain which actions are good. As Moore acknowledged, the open question argument does not do anything to show that pleasure, for example, is not the sole criterion of the goodness of an action. It shows only that this cannot be known to be true by definition, and so, if it is to be known at all, it must be known by some other means.

In spite of these doubts, Moore's argument was widely accepted at the time as showing that all attempts to derive ethical conclusions from anything not itself ethical in nature are bound to fail. The point was soon seen to be related to that made by Hume in his remarks on writers who move from "is" to "ought." Moore, however, would have considered Hume's own account of morality to be naturalistic because of its definition of virtue in terms of the sentiments of the spectator. The upshot was that for 30 years after the publication of *Principia Ethica* intuitionism was the dominant metaethical position in British philosophy. In addition to Moore, its supporters included H.A. Prichard and Sir W.D. Ross.

Modern intuitionism. The 20th-century intuitionists were not far removed philosophically from their 18th-century predecessors—those such as Richard Price who had learned from Hume's criticism and did not attempt to reason his way to ethical conclusions but claimed rather that ethical knowledge is gained through an immediate apprehension of its truth. In other words, a true ethical judgment is self-evident as long as we are reflecting clearly and calmly and our judgment is not distorted by self-interest or faulty moral upbringing. Ross, for example, took "the convictions of thoughtful, well-educated people" as "the data of ethics," observing that while some may be illusory, they should only be rejected when they conflict with others that are better able to stand up to "the test of reflection."

The intuitionists differed on the nature of the moral truths that are apprehended in this way. For Moore it was self-evident that certain things are valuable: *e.g.*, the pleasures of friendship and the enjoyment of beauty. On the other hand, Ross thought we know it to be our duty to do acts of a certain type. These differences will be dealt with in the discussion of normative ethics. They are, however, significant to metaethical intuitionism because they reveal the lack of agreement, even among the intuitionists themselves, about moral judgments that each claims to be self-evident.

This disagreement was one of the reasons for the eventual rejection of intuitionism, which, when it came, was as complete as its acceptance had been in earlier decades. But there was also a more powerful philosophical motive working against intuitionism. During the 1930s, Logical Positivism, brought from Vienna by Ludwig Wittgenstein and popularized by A.J. Ayer in his manifesto *Language, Truth and Logic* (1936), became influential in British philosophy. According to the Logical Positivists, all true statements fall into two categories: logical truths and statements of fact. Moral judgments cannot fit comfortably into either category. They cannot be logical truths, for these are mere tautologies that can tell us nothing more than what is already contained in the definitions of the terms. Nor can they be statements of fact because these must, according to the Logical Positivists, be at least in principle verifiable; there is no way of verifying the truths that the intuitionists claimed to apprehend. The truths of mathematics, on which intuitionists had continued to rely as the one clear parallel case of a truth known by its self-evidence, were explained now as logical truths. In this view, mathematics tells us nothing about the world; it is simply a logical system, true by the definitions of the terms involved, which may be useful in our dealings with the world. Thus the intuitionists lost the one useful analogy to which they could appeal in support of the existence of a body of self-evident truths known by reason alone. It seemed to follow that moral judgments could not be truths at all.

Impact of Logical Positivism

Emotivism. In his above-cited *Language, Truth and Logic,* Ayer offered an alternative account: moral judgments are not statements at all. When we say, "You acted wrongly in stealing that money," we are not expressing any fact beyond that stated by "You stole that money." It is, however, as if we had stated this fact with a special tone of abhorrence, for in saying that something is wrong, we are expressing our feelings of disapproval toward it.

This view was more fully developed by Charles Stevenson in *Ethics and Language* (1945). As the titles of books of this period suggest, philosophers were now paying more attention to language and to the different ways in which it could be used. Stevenson distinguished the facts a sentence may convey from the emotive impact it is intended to have. Moral judgments are significant, he urged, because of their emotive impact. In saying that something is wrong, we are not merely expressing our disapproval of it, as Ayer suggested. We are encouraging those to whom we speak to share our attitude. This is why we bother to

argue about our moral views, while on matters of taste we may simply agree to differ. It is important to us that others share our attitudes on war, equality, or killing; we do not care if they prefer to take their tea with lemon and we do not.

Charges of subjectivism

The emotivists were immediately accused of being subjectivists. In one sense of the term subjectivist, the emotivists could firmly reject this charge. Unlike other subjectivists in the past, they did not hold that those who say, for example, "Stealing is wrong," are making a statement of fact about their own feelings or attitudes toward stealing. This view—more properly known as subjective naturalism because it makes the truth of moral judgments depend on a natural, albeit subjective, fact about the world—could be refuted by Moore's open question argument. It makes sense to ask: "I know that I have a feeling of approval toward this, but is it good?" It was the emotivists' view, however, that moral judgments make no statements of fact at all. The emotivists could not be defeated by the open question argument because they agreed that no definition of "good" in terms of facts, natural or unnatural, could capture the emotive element of its meaning. Yet, this reply fails to confront the real misgivings behind the charge of subjectivism: the concern that there are no possible standards of right and wrong other than one's own subjective feelings. In this sense, the emotivists were subjectivists.

Existentialism. About this time a different form of subjectivism was becoming fashionable on the Continent and to some extent in the United States. Existentialism was as much a literary as a philosophical movement. Its leading figure, Jean-Paul Sartre, propounded his ideas in novels and plays as well as in his major philosophical treatise, *L'Être et le néant* (1943; *Being and Nothingness*). For Sartre, because there is no God, human beings have not been designed for any particular purpose. The Existentialists express this by stating that our existence precedes our essence. In saying this, they make clear their rejection of the Aristotelian notion that just as we can recognize a good knife once we know that the essence of a knife is to cut, so we can recognize a good human being once we understand the essence of human nature. Because we have not been designed for any specific end, we are free to choose our own essence, which means to choose how we will live. To say that we are compelled by our situation, our nature, or our role in life to act in a certain way is to exhibit "bad faith." This seems to be the only term of disapproval the Existentialists are prepared to use. As long as we choose "authentically," there are no moral standards by which our conduct can be criticized.

Relativity of all moral standards

This, at least, is the view most widely held by the Existentialists. In one work, a brochure entitled *L'Existentialisme est un humanisme* (1946; "Existentialism Is a Humanism"; Eng. trans., *Existentialism and Humanism*), Sartre backs away from so radical a subjectivism by suggesting a version of Kant's idea that we must be prepared to apply our judgments universally. He does not reconcile this view with conflicting statements elsewhere in his writings, and it is doubtful if it can be regarded as a statement of his true ethical views. It may reflect, however, a widespread postwar reaction to the spreading knowledge of what happened at Auschwitz and other Nazi death camps. One leading German prewar Existentialist, Martin Heidegger, had actually become a Nazi. Was this "authentic choice" just as good as Sartre's own choice to join the French Résistance? Is there really no firm ground from which such a choice could be rejected? This seemed to be the upshot of the pure Existentialist position, just as it was an implication of the ethical emotivism that was dominant among English-speaking philosophers. It is scarcely surprising that many philosophers should search for a metaethical view that did not commit them to this conclusion. The means used by Sartre in *L'Existentialisme est un humanisme* were also to have their parallel, though in a much more sophisticated form, in British moral philosophy.

Universal prescriptivism. In *The Language of Morals* (1952), R.M. Hare supported some of the elements of emotivism but rejected others. He agreed that in making moral judgments we are not primarily seeking to describe anything; but neither, he said, are we simply expressing our attitudes. Instead, he suggested that moral judgments prescribe; that is, they are a form of imperative sentence. Hume's rule about not deriving an "is" from an "ought" can best be explained, according to Hare, in terms of the impossibility of deriving any prescription from a set of descriptive sentences. Even the description "There is an enraged bull bearing down on you" does not necessarily entail the prescription "Run!" because I may have been searching for ways of killing myself in such a way that my children can still benefit from my life insurance. Only I can choose whether the prescription fits what I want. Herein lies moral freedom: because the choice of prescription is individual, no one can tell another what he or she must think right.

Hare's espousal of the view that moral judgments are prescriptions led commentators on his first book to classify him with the emotivists as one who did not believe in the possibility of using reason to arrive at ethical conclusions. That this was a mistake became apparent with the publication of his second book, *Freedom and Reason* (1963). The aim of the book was to show that the moral freedom guaranteed by prescriptivism is, notwithstanding its element of choice, compatible with a substantial amount of reasoning about moral judgments. Such reasoning is possible, Hare wrote, because moral judgments must be "universalizable." This notion owed something to the ancient Golden Rule and even more to Kant's first formulation of the categorical imperative. In Hare's treatment, however, these ideas were refined so as to eliminate their obvious defects. Moreover, for Hare universalizability is not a substantive moral principle but a logical feature of the moral terms. This means that anyone who uses such terms as right and ought is logically committed to universalizability.

Definition of universalizability

To say that a moral judgment must be universalizable means, for Hare, that if I judge a particular action—say, a man's embezzlement of a million dollars from his employer—to be wrong, I must also judge any relevantly similar action to be wrong. Of course, everything will depend on what is allowed to count as a relevant difference. Hare's answer is that all features may count, except those that contain ineliminable uses of words such as I or my, or singular terms such as proper names. In other words, the fact that he embezzled a million dollars in order to be able to take holidays in Tahiti, whereas I embezzled the same sum so as to channel it from my wealthy employer to those starving in Africa, may be a relevant difference; the fact that the man's crime benefitted him, whereas my crime benefitted me, cannot be so.

This notion of universalizability can also be used to test whether a difference that is alleged to be relevant—for instance, skin colour or even the position of a freckle on one's nose—really is relevant. Hare emphasized that the same judgment must be made in all conceivable cases. Thus if a Nazi were to claim that he may kill a person because that person is Jewish, he must be prepared to prescribe that if, somehow, it should turn out that he is himself of Jewish origin, he should also be killed. Nothing turns on the likelihood of such a discovery; the same prescription has to be made in all hypothetically, as well as actually, similar cases. Since only an unusually fanatical Nazi would be prepared to do this, universalizability is a powerful means of reasoning against certain moral judgments, including those made by the Nazis. At the same time, since there could be fanatical Nazis who are prepared to die for the purity of the Aryan race, the argument of *Freedom and Reason* allows that the role played by reason in ethics does have definite limits. Hare's position at this stage, therefore, appeared to be a compromise between the extreme subjectivism of the emotivists and some more objectivist view of ethics. As so often happens with those who try to take the middle ground, Hare was soon to receive criticism from both sides.

Modern naturalism. For a time, Moore's presentation of the naturalistic fallacy halted attempts to define "good" in terms of natural qualities such as happiness. The effect was, however, both local and temporary. In the United States, Ralph Barton Perry was untroubled by Moore's arguments. His *General Theory of Value* (1926) gave an account of value that was objectivist and much less mys-

terious than the intuitionist accounts, which were at that time dominating British philosophy. Perry suggested that there is no such thing as value until a being desires something, and nothing can have intrinsic value considered apart from all desiring beings. A novel, for example, has no value at all unless there is a being who desires to read it or perhaps use it for some other purpose, such as starting a fire on a cold night. Thus Perry is a naturalist, for he defines value in terms of the natural quality of being desired or, as he puts it, being an object of an interest. His naturalism is objectivist, in spite of this dependence of value on desires, because value is defined as any object of any interest. Accordingly, even if I do not desire, say, this encyclopaedia for any purpose at all, I cannot deny that it has some value so long as there is some being who does desire it. Moreover, Perry believed it followed from his theory that the greatest moral value is to be found in whatever leads to the harmonious integration of interests.

In Britain, Moore's impact was for a long time too great for any form of naturalism to be taken seriously. It was only as a response to Hare's intimation that any principle could be a moral principle so long as it satisfied the formal requirement of universalizability that philosophers such as Philippa Foot, Elizabeth Anscombe, and Geoffrey Warnock began to suggest that perhaps a moral principle must also have a particular kind of content—*i.e.,* it must deal, for instance, with some aspect of wants, welfare, or flourishing.

The problem with these suggestions, Hare soon pointed out, is that if we define morality in such a way that moral principles are restricted to those that maximize well-being, then if there is a person who is not interested in maximizing well-being, moral principles, as we have defined them, will have no prescriptive force for that person. This reply elicited two responses—namely, those of Anscombe and Foot.

Anscombe went back to Aristotle, suggesting that we need a theory of human flourishing that will provide an account of what any person must do in order to flourish, and so will lead to a morality that every one of us has reason to follow. No such theory was forthcoming, however, until 1980 when John Finnis offered a theory of basic human goods in his *Natural Law and Natural Rights.* The book was acclaimed by Roman Catholic moral theologians and philosophers, but natural law ethics continues to have few followers outside these circles.

Foot initially attempted to defend a similarly Aristotelian view in which virtue and self-interest are necessarily linked, but she came to the conclusion that this link could not be made. This led her to abandon the assumption that we all have adequate reasons for doing what is right. Like Hume, she suggested that it depends on what we desire and especially on how much we care about others. She observed that morality is a system of hypothetical, not categorical, imperatives.

A much cruder form of naturalism surfaced from a different direction with the publication of Edward O. Wilson's *Sociobiology: The New Synthesis* (1975). Wilson, a biologist rather than a philosopher, claimed that new developments in the application of evolutionary theory to social behaviour would allow ethics to be "removed from the hands of philosophers" and "biologicized." It was not the first time that a scientist, frustrated by the apparent lack of progress in ethics as compared to the sciences, had proposed some way of transforming ethics into a science. In a later book, *On Human Nature* (1978), Wilson suggested that biology justifies specific values (including the survival of the gene pool) and, because man is a mammal rather than a social insect, universal human rights. Other sociobiologists have gone further still, reviving the claims of earlier "social Darwinists" to the effect that Darwin's theory of evolution shows why it is right that there should be social inequality.

Attempt to transform ethics into a science

As the above section on the origin of ethics suggests, evolutionary theory may indeed have something to reveal about the origins and nature of the systems of morality used by human societies. Wilson is, however, plainly guilty of breaching Hume's rule when he tries to draw from a theory of a factual nature ethical premises that tell us what we ought to do. It may be that, coupled with the premise that we wish our species to survive for as long as possible, evolutionary theory will suggest the direction we ought to take, but even that premise cannot be regarded as unquestionable. It is not impossible to imagine circumstances in which life is so grim that extinction is preferable. That choice cannot be dictated by science. It is even less plausible to suppose that more specific choices about social equality can be settled by evolutionary theory. At best, the theory would indicate the costs we might incur by moving to greater equality; it could not conceivably tell us whether incurring those costs is justifiable.

Recent developments in metaethics. In view of the heat of the debate between Hare and his naturalist opponents during the 1960s, the next development was surprising. At first in articles and then in the book *Moral Thinking* (1981), Hare offered a new understanding of what is involved in universalizability that relies on treating moral ideals in a similar fashion to ordinary desires or preferences. In *Freedom and Reason* the universalizability of moral judgments prevented me from giving greater weight to my own interests, simply on the grounds that they are mine, than I was prepared to give to anyone else's interests. In *Moral Thinking* Hare argued that to hold an ideal, whether it be a Nazi ideal such as the purity of the Aryan race or a more conventional ideal such as that justice must be done irrespective of the consequences, is really to have a special kind of preference. When I ask whether I can prescribe a moral judgment universally, I must take into account all the ideals and preferences held by all those who will be affected by the action I am judging; and in taking these into account, I cannot give any special weight to my own ideals merely because they are my own. The effect of this application of universalizability is that for a moral judgment to be universalizable it must ultimately be based on the maximum possible satisfaction of the preferences of all those affected by it. Thus Hare claimed that his reading of the formal property of universalizability inherent in moral language enables him to solve the ancient problem of showing how reason can, at least in principle, resolve ethical disagreement. Moral freedom, on the other hand, has been reduced to the freedom to be an amoralist and to avoid using moral language altogether.

Hare's position was immediately challenged by J.L. Mackie in *Ethics: Inventing Right and Wrong* (1977). In the course of a defense of moral subjectivism, Mackie argued that Hare had stretched the notion of universalizability far beyond anything that is really inherent in moral language. Moreover, even if such a notion were embodied in our way of thinking and talking about morality, Mackie insisted that we would always be free to reject such notions and to decide what to do without concerning ourselves with whether our judgments are universalizable in Hare's, or indeed in any, sense. According to Mackie, our ordinary use of moral language presupposes that moral judgments are statements about something in the universe and, therefore, can be true or false. This is, however, a mistake. Drawing on Hume, Mackie says that there cannot be any matters of fact that make it rational for everyone to act in a certain way. If we do not reject morality altogether, we can only base our moral judgments on our own desires and feelings.

There are a number of contemporary British philosophers who do not accept either Hare's or Mackie's metaethical views. Those who hold forms of naturalism have already been mentioned. Others, including the Oxford philosophers David Wiggins and John McDowell, have employed modern semantic theories of the nature of truth to show that even if moral judgments do not correspond to any objective facts or self-evident truths, they may still be proper candidates for being true or false. This position has become known as moral realism. For some, it makes moral judgments true or false at the cost of taking objectivity out of the notion of truth.

Moral realism

Many modern writers on ethics, including Mackie and Hare, share a view of the nature of practical reason derived from Hume. Our reasons for acting morally, they hold, must depend on our desires because reason in action applies only to the best way of achieving what we desire.

This view of practical reason virtually precludes any general answer to the question "Why should I be moral?" Until very recently, this question had received less attention in the 20th century than in earlier periods. In the early part of the century, such intuitionists as H.A. Prichard had rejected all attempts to offer extraneous reasons for being moral. Those who understood morality would, they said, see that it carried its own internal reasons for being followed. For those who could not see these reasons, the situation was reminiscent of the story of the emperor's new clothes.

The question fared no better with the emotivists. They defined morality so broadly that anything an individual desires can be considered to be moral. Thus there can be no conflict between morality and self-interest, and if anyone asks "Why should I be moral?" the emotivist response would be to say "Because whatever you most approve of doing is, by definition, your morality." Here the question is effectively being rejected as senseless, but this reply does nothing to persuade the questioners to act in a benevolent or socially desirable way. It merely tells them that no matter how antisocial their actions may be, they can still be moral as the emotivists define the term.

For Hare, on the other hand, the question "Why should I be moral?" amounts to asking why I should act only on those judgments that I am prepared to universalize; and the answer he gives is that unless this is what I want to do, it is not always possible to give an adult a reason for doing so. At the same time, Hare does believe that if someone asks why children should be brought up to be morally good, the answer is that they are more likely to be happy if they develop habits of acting morally.

Other philosophers have put the question to one side, saying that it is a matter for psychologists rather than for philosophers. In earlier periods, of course, psychology was considered a branch of philosophy rather than a separate discipline, but in fact psychologists have also had little to say about the connection between morality and self-interest. In *Motivation and Personality* (1954) and other works, Abraham H. Maslow developed a psychological theory reminiscent of Shaftesbury in its optimism about the link between personal happiness and moral values, but Maslow's factual evidence was thin. Victor Emil Frankl, a psychotherapist, has written several popular books defending a position essentially similar to that of Joseph Butler on the attainment of happiness. The gist of this view is known as the paradox of hedonism. In *The Will to Meaning* (1969), Frankl states that those who aim directly at happiness do not find it; those whose lives have meaning or purpose apart from their own happiness find happiness as well.

Psychological theories on the connection between morality and self-interest

The U.S. philosopher Thomas Nagel has taken a different approach to the question of how we may be motivated to act altruistically. Nagel challenges the assumption that Hume was right about reason being subordinate to desires. In *The Possibility of Altruism* (1969), Nagel sought to show that if reason must always be based on desire, even our normal idea of prudence (that we should give the same weight to our future pains and pleasures as we give to our present ones) becomes incoherent. Once we accept the rationality of prudence, however, Nagel argued that a very similar line of argument can lead us to accept the rationality of altruism—*i.e.*, the idea that the pains and pleasures of another individual are just as much a reason for one to act as are one's own pains and pleasures. This means that reason alone is capable of motivating moral action; hence, it is unnecessary to appeal to self-interest or benevolent feelings. Though not an intuitionist in the ordinary sense, Nagel has effectively reopened the 18th-century debate between the moral sense school and the intuitionists who believed that reason alone can play a role in action.

The most influential work in ethics by a U.S. philosopher since the early 1960s, John Rawls's *Theory of Justice* (1971), is for the most part centred on normative ethics, and so will be discussed in the next section; it has, however, had some impact in metaethics as well. To argue for his principles of justice, Rawls uses the idea of a hypothetical contract, in which the contracting parties are behind a "veil of ignorance" that prevent them from knowing any particular details about their own attributes. Thus one cannot try to benefit oneself by choosing principles of justice that favour the wealthy, the intelligent, males, or whites. The effect of this requirement is in many ways similar to Hare's idea of universalizability, but Rawls claims that it avoids, as the former does not, the trap of grouping together the interests of different individuals as if they all belonged to one person. Accordingly, the old social contract model that had largely been neglected since the time of Rousseau has had a new wave of popularity as a form of argument in ethics.

Revival of the notion of social contract

The other aspect of Rawls's thought to have metaethical significance is his so-called method of reflective equilibrium—the idea that a sound moral theory is one that matches reflective moral judgments. In *A Theory of Justice* Rawls uses this method to justify tinkering with the original model of the hypothetical contract until it produces results that are not too much at odds with ordinary ideas of justice. To his critics, this represents a reemergence of a conservative form of intuitionism, for it means that new moral theories are tested against ordinary moral intuitions. If a theory fails to match enough of these, it will be rejected no matter how strong its own foundations may be. In Rawls's defense it may be said that it is only our "reflective moral judgments" that serve as the testing ground—our ordinary moral intuitions may be rejected, perhaps simply because they are contrary to a well-grounded theory. If such be the case, the charge of conservatism may be misplaced, but in the process the notion of some independent standard by which the moral theory may be tested has been weakened, perhaps so far as to become virtually meaningless.

Perhaps the most impressive work of metaethics published in the United States in recent years is R.B. Brandt's *Theory of the Good and the Right* (1979). Brandt returns to something like the naturalism of Ralph Barton Perry but with a distinctive late 20th-century American twist. He spends little time on the concept of good, believing that everything capable of being expressed by this word can be more clearly stated in terms of rational desires. To explicate this notion of a rational desire, Brandt appeals to cognitive psychotherapy. An ideal process of cognitive psychotherapy would eliminate many desires: those based on false beliefs, those which one has only because one is ignoring the feelings or desires that are likely to be expressed in the future, the desires or aversions that are artificially caused by others, desires that are based on early deprivation, and so on. The desires that an individual would still have, undiminished in strength after going through this process, are what Brandt is prepared to call rational desires.

Brandt's notion of rational desire

In contrast to his view of the term good, Brandt does think that the notions of morally right and morally wrong are useful. He suggests that, in calling an action morally wrong, we should mean that it would be prohibited by any moral code that all fully rational people would support for the society in which they are to live. (Brandt then argues that fully rational people would support that moral code which would maximize happiness, but the justification of this claim is a task for normative ethics, not metaethics.)

Brandt's final chapter is an indication of the revival of interest in the question, as he phrases it, "Is it always rational to act morally?" His answer, echoing Shaftesbury in modern guise, is that such desires as benevolence would survive cognitive psychotherapy, and so a rational person would be benevolent. A rational person would also have other moral motives, including an aversion to dishonesty. These motives will occasionally conflict with self-interested desires, and there can be no guarantee that the moral motives will be the stronger. If they are not, and in spite of the fact that a rational person would support a code favouring honesty, Brandt is unable to say that it would be irrational to follow self-interest rather than morality. A fully rational person might support a certain kind of moral code and yet not act in accordance with it on every occasion.

As the century draws to a close, the issues that divided Plato and the Sophists are still dividing moral philoso-

phers. Ironically, the one position that now has few defenders is Plato's view that "good" refers to an idea or property having an objective existence quite apart from anyone's attitudes or desires—on this point the Sophists appear to have won out at last. Yet, this still leaves ample room for disagreement about the extent to which reason can bring about agreed decisions on what we ought to do. There also remains the dispute about whether it is proper to refer to moral judgments as true and false. On the other central question of metaethics, the relationship between morality and self-interest, a complete reconciliation of the two continues to prove—at least for those not prepared to appeal to a belief in reward and punishment in another life—as elusive as it did for Sidgwick at the end of the 19th century.

NORMATIVE ETHICS

The debate over consequentialism. Normative ethics seeks to set norms or standards for conduct. The term is commonly used in reference to the discussion of general theories about what one ought to do, a central part of Western ethics since ancient times. Normative ethics continued to hold the spotlight during the early years of the 20th century, with intuitionists such as W.D. Ross engaged in showing that an ethic based on a number of independent duties was superior to Utilitarianism. With the rise of Logical Positivism and emotivism, however, the logical status of normative ethics seemed doubtful: Was it not simply a matter of whatever one approved? Nor was the analysis of language, which dominated philosophy in English-speaking countries during the 1950s, any more congenial to normative ethics. If philosophy could do no more than analyze words and concepts, how could it offer guidance about what one ought to do? The subject was therefore largely neglected until the 1960s, when emotivism and linguistic analysis were both on the retreat and moral philosophers once again began to think about how individuals ought to live.

A crucial question of normative ethics is whether actions are to be judged right or wrong solely on the basis of their consequences. Traditionally, those theories that judge actions by their consequences have been known as teleological theories, while those that judge actions according to whether they fall under a rule have been referred to as deontological theories. Although the latter term continues to be used, the former has been replaced to a large extent by the more straightforward term consequentialist. The debate over this issue has led to the development of different forms of consequentialist theories and to a number of rival views.

Classical Utilitarianism is the simplest form

Varieties of consequentialism. The simplest form of consequentialism is classical Utilitarianism, which holds that every action is to be judged good or bad according to whether its consequences do more than any alternative action to increase—or, if that is impossible, to limit any unavoidable decrease in—the net balance of pleasure over pain in the universe. This is often called hedonistic Utilitarianism.

G.E. Moore's normative position offers an example of a different form of consequentialism. In the final chapters of the aforementioned *Principia Ethica* and also in *Ethics* (1912), Moore argued that the consequences of actions are decisive for their morality, but he did not accept the classical Utilitarian view that pleasure and pain are the only consequences that matter. Moore asked his readers to picture a world filled with all possible imaginable beauty but devoid of any being who can experience pleasure or pain. Then the reader is to imagine another world, as ugly as can be but equally lacking in any being who experiences pleasure or pain. Would it not be better, Moore asked, that the beautiful world rather than the ugly world exist? He was clear in his own mind that the answer was affirmative, and he took this as evidence that beauty is good in itself, apart from the pleasure it brings. He also considered that the friendship of close personal relationships has a similar intrinsic value independent of its pleasantness. Moore thus judged actions by their consequences but not solely by the amount of pleasure they produced. Such a position was once called ideal Utilitarianism because it was a form of Utilitarianism based on certain ideals. Today, however, it is more frequently referred to by the general label consequentialism, which includes, but is not limited to, Utilitarianism.

R.M. Hare is another example of a consequentialist. His interpretation of universalizability leads him to the view that for a judgment to be universalizable, it must prescribe what is most in accord with the preferences of all those affected by the action. This form of consequentialism is frequently called preference Utilitarianism because it attempts to maximize the satisfaction of preferences, just as classical Utilitarianism endeavours to maximize pleasure or happiness. Part of the attraction of such a view lies in the way in which it avoids making judgments about what is intrinsically good, finding its content instead in the desires that people, or sentient beings generally, do have. Another advantage is that it overcomes the objection, which so deeply troubled Mill, that the production of simple, mindless pleasure becomes the supreme goal of all human activity. Against these advantages we must put the fact that most preference Utilitarians want to base their judgments, not on the desires that people actually have, but rather on those they would have if they were fully informed and thinking clearly. It then becomes essential to discover what people would want under these conditions, and, because most people most of the time are less than fully informed and clear in their thoughts, the task is not an easy one.

Preference Utilitarianism

It may also be noted in passing that Hare claims to derive his version of Utilitarianism from universalizability, which in turn he draws from moral language and moral concepts. Moore, on the other hand, had simply found it self-evident that certain things were intrinsically good. Another Utilitarian, the Australian philosopher J.J.C. Smart, has defended hedonistic Utilitarianism by asserting that he has a favourable attitude to making the surplus of happiness over misery as large as possible. As these differences suggest, consequentialism can be held on the basis of widely differing metaethical views.

Consequentialists may also be separated into those who ask of each individual action whether it will have the best consequences, and those who ask this question only of rules or broad principles and then judge individual actions by whether they fall under a good rule or principle. The distinction having arisen in the specific context of Utilitarian ethics, the former are known as act-Utilitarians and the latter as rule-Utilitarians.

Rule-Utilitarianism developed as a means of making the implications of Utilitarianism less shocking to ordinary moral consciousness. (The germ of this approach is seen in Mill's defense of Utilitarianism.) There might be occasions, for example, when stealing from one's wealthy employer in order to give to the poor would have good consequences. Yet, surely it would be wrong to do so. The rule-Utilitarian solution is to point out that a general rule against stealing is justified on Utilitarian grounds, because otherwise there could be no security of property. Once the general rule has been justified, individual acts of stealing can then be condemned whatever their consequences because they violate a justifiable rule.

This suggests an obvious question, one already raised by the above account of Kant's ethics: How specific may the rule be? Although a rule prohibiting stealing may have better consequences than no rule at all against stealing, would not the best consequences of all follow from a rule that permitted stealing only in those special cases in which it is clear that stealing will have better consequences than not stealing? But what then is the difference between act- and rule-Utilitarianism? In *Forms and Limits of Utilitarianism* (1965), David Lyons argued that if the rule were formulated with sufficient precision to take into account all its causally relevant consequences, rule-Utilitarianism would collapse into act-Utilitarianism. If rule-Utilitarianism is to be maintained as a distinct position, then there must be some restriction on how specific the rule can be so that at least some relevant consequences are not taken into account.

To ignore relevant consequences is to break with the very essence of consequentialism; rule-Utilitarianism is

Rule-Utilitarianism versus act-Utilitarianism

therefore not a true form of Utilitarianism at all. That, at least, is the view taken by Smart, who has derided rule-Utilitarianism as "rule-worship" and consistently defended act-Utilitarianism. Of course, when time and circumstances make it awkward to calculate the precise consequences of an action, Smart's act-Utilitarian will resort to rough and ready "rules of thumb" for guidance; but these rules of thumb have no independent status apart from their usefulness in predicting likely consequences, and if ever we are clear that we will produce better consequences by acting contrary to the rule of thumb, we should do so. If this leads us to do things that are contrary to the rules of conventional morality, then, Smart says, so much the worse for conventional morality.

Today, straightforward rule-Utilitarianism has few supporters. On the other hand, a number of more complex positions have been proposed, bridging in some way the distance between rule-Utilitarianism and act-Utilitarianism.

Intuitive level of moral thought

In *Moral Thinking* Hare distinguished two levels of thought about what we ought to do. At the critical level we may reason about the principles that should govern our action and consider what would be for the best in a variety of hypothetical cases. The correct answer here, Hare believed, is always that the best action will be the one that has the best consequences. This principle of critical thinking is not, however, well-suited for everyday moral decision making. It requires calculations that are difficult to carry out under the most ideal circumstances and virtually impossible to carry out properly when we are hurried or liable to be swayed by our emotions or our interests. Everyday moral decisions are the proper domain of the intuitive level of moral thought. At this intuitive level we do not enter into fine calculations of consequences; instead, we act in accordance with fundamental moral principles that we have learned and accepted as determining, for practical purposes, whether an act is right or wrong. Just what these moral principles should be is a task for critical thinking. They must be the principles that, when applied intuitively by most people, will produce the best consequences overall, and they must also be sufficiently clear and brief to be made part of the moral education of children. Hare therefore can avoid the dilemma of the rule-Utilitarian while still preserving the advantages of that position. Given that ordinary moral beliefs reflect the experience of many generations, Hare believed that judgments made at the intuitive level will probably not be too different from judgments made by conventional morality. At the same time, Hare's restriction on the complexity of the intuitive principles is fully consequentialist in spirit.

Some recently published work has gone further still in this direction. Following on earlier discussions of the difficulties consequentialists may have in trusting one another—since the word of a Utilitarian is only as good as the consequences of keeping the promise appear to him to be—Donald Regan has explored the problems of cooperation among Utilitarians in his *Utilitarianism and Co-operation* (1980) and has come out with a further variation designed to make cooperation feasible and thus to achieve the best consequences on the whole. In *Reasons and Persons* (1984), Derek Parfit argued that to aim always at producing the best consequences would be indirectly self-defeating; we would be cutting ourselves off from some of the greatest goods of human life, including those close personal relationships that demand that we sacrifice the ideal of impartial benevolence to all in order that we may give preference to those we love. We therefore need, Parfit suggested, not simply a theory of what we should all do, but a theory of what motives we should all have. Parfit, like Hare, plausibly contended that recognizing this distinction will bring the practical application of consequentialist theories closer to conventional moral judgments.

An ethic of prima facie duties. In the first third of the 20th century, it was the intuitionists, especially W.D. Ross, who provided the major alternative to Utilitarianism. Because of this situation, the position described below is sometimes called intuitionism, but it seems less likely to cause confusion if we reserve that label for the quite distinct metaethical position held by Ross—and incidentally by Sidgwick as well—and refer to the normative position by the more descriptive label, an "ethic of prima facie duties."

Ross's normative ethic consists of a list of duties, each of which is to be given independent weight: fidelity, reparation, gratitude, beneficence, nonmaleficence, and self-improvement. If an act falls under one and only one of these duties, it ought to be carried out. Often, of course, an act will fall under two or more duties: I may owe a debt of gratitude to someone who once helped me, but beneficence will be better served if I help others in greater need. This is why the duties are, Ross says, prima facie rather than absolute; each duty can be overridden if it conflicts with a more stringent duty.

An ethic structured in this manner may match our ordinary moral judgments more closely than a consequentialist ethic, but it suffers from two serious drawbacks. First, how can we be sure that just those duties listed by Ross are independent sources of moral obligation? Ross could only respond that if we examine them closely we will find that these, and these alone, are self-evident. But others, even other intuitionists, have found that what was self-evident to Ross was not self-evident to them. Second, if we grant Ross his list of independent prima facie moral duties, we still need to know how to decide, in a particular situation, when a less stringent duty is overridden by a more stringent one. Here, too, Ross had no better answer than an unsatisfactory appeal to intuition.

Rawls's theory of justice. When philosophers again began to take an interest in normative ethics in the 1960s after an interval of some 30 years, no theory could rival the ability of Utilitarianism to provide a plausible and systematic basis for moral judgments in all circumstances. Yet, many people found themselves unable to accept Utilitarianism. One common ground for dissatisfaction was that Utilitarianism does not offer any principle of justice beyond the basic idea that everyone's happiness—or preferences, depending on the form of Utilitarianism—counts equally. Such a principle is quite compatible with sacrificing the welfare of some to the greater welfare of others. This situation explains the enthusiastic welcome accorded to Rawls's *Theory of Justice* when it appeared in 1971.

Rawls's theory as an alternative to Utilitarianism

Rawls offered an alternative to Utilitarianism that came close to matching its rival's ability to provide a systematic theory of what one ought to do and, at the same time, led to conclusions about justice very different from those of the Utilitarians.

Rawls asserted that if people had to choose principles of justice from behind a "veil of ignorance" that restricted what they could know of their own position in society, they would not seek to maximize overall utility. Instead, they would safeguard themselves against the worst possible outcome, first, by insisting on the maximum amount of liberty compatible with the like liberty for others; and, second, by requiring that wealth be distributed so as to make the worst-off members of the society as well-off as possible. This second principle is known as the "maximin" principle, because it seeks to maximize the welfare of those at the minimum level of society. Such a principle might be thought to lead directly to an insistence on the equal distribution of wealth, but Rawls points out that if we accept certain assumptions about the effect of incentives and the benefits that may flow to all from the productive labours of the most talented members of society, the maximin principle could allow considerable inequality.

The "maximin" principle

In the decade following its appearance, *A Theory of Justice* was subjected to unprecedented scrutiny by moral philosophers throughout the world. Two major issues emerged: Were the two principles of justice soundly derived from the original contract situation? And did the two principles amount, in themselves, to an acceptable theory of justice?

To the first question, the general verdict was negative. Without appealing to specific psychological assumptions about an aversion to risk—and Rawls disclaimed any such assumptions—there was no convincing way in which Rawls could exclude the possibility that the parties to the original contract would choose to maximize average

utility, thus giving themselves the best possible chance of having a high level of welfare. True, each individual making such a choice would have to accept the possibility that he would end up with a very low level of welfare, but that might be a risk worth running for the sake of a chance at a very high level.

Even if the two principles cannot validly be derived from the original contract, they might be sufficiently attractive to stand on their own either as self-evident moral truths—if we are objectivists—or as principles to which we might have favourable attitudes. Maximin, in particular, has proved attractive in a variety of disciplines, including welfare economics, a field in which preference Utilitarianism once reigned unchallenged. But maximin has also had its critics, who have pointed out that the principle could require us to forgo very great benefits to the vast majority if, for some reason, this would require some loss (no matter how trivial) to the worst-off members of society.

Rights theories. One of Rawls's severest critics, Robert Nozick of the United States, rejected the assumption that lies behind not only the maximin principle but behind any principle that seeks to achieve a pattern of distribution by taking from one group in order to give to another. In attempting to bring about a certain pattern of distribution, Nozick said, these principles ignore the question of how the individuals from whom wealth will be taken acquired their wealth in the first place. If they have done so by wholly legitimate means without violating the rights of others, then Nozick held that no one, not even the state, can have the right to take their wealth from them without their consent.

Although appeals to rights have been common since the great 18th-century declarations of the rights of man, most ethical theorists have treated rights as something that must be derived from more basic ethical principles or else from accepted social and legal practices. Recently, however, there have been attempts to turn this tendency around and make rights the basis of the ethical theory. It is in the United States, no doubt because of its history and constitution, that the appeal to rights as a fundamental moral principle has been most common. Nozick's *Anarchy, State and Utopia* (1974) is one example of a rights-based theory, although it is mostly concerned with the application of the theory in the political sphere and says very little about other areas of normative ethics. Unlike Rawls, who for all his disagreement with Utilitarianism is still a consequentialist of sorts, Nozick is a deontologist. Our rights to life, liberty, and legitimately acquired property are absolute, and no act can be justified if it violates them. On the other hand, we have no duty to assist people in the preservation of their rights. If others go about their own affairs without infringing on the rights of others, I must not infringe on their rights; but if they are starving, I have no duty to share my food with them. We can appeal to the generosity of the rich, but we have absolutely no right to tax them against their will so as to provide relief for the poor. This doctrine has found favour with some Americans on the political right, but it has proved too harsh for most students of ethics.

Rights as the basis of ethical theory

To illustrate the variety of possible theories based on rights, we can take as another example the one propounded by Ronald Dworkin in *Taking Rights Seriously* (1977). Dworkin agreed with Nozick that rights are not to be overridden for the sake of improved welfare: rights are, he said, "trumps" over ordinary consequentialist considerations. Dworkin's view of rights, however, derives from a fundamental right to equal concern and respect. This makes it much broader than Nozick's theory, since respect for others may require us to assist them and not merely leave them to fend for themselves. Accordingly, Dworkin's view obliges the state to intervene in many areas to ensure that rights are respected.

In its emphasis on equal concern and respect, Dworkin's theory is part of a recent revival of interest in Kant's principle of respect for persons as the fundamental principle of ethics. This principle, like the principle of justice, is often said to be ignored by Utilitarians. Rawls invoked it when setting out the underlying rationale of his theory of justice. The concept, however, suffers from vagueness, and attempts to develop it into something more specific that could serve as the basis for a complete ethical theory have not—unless Rawls's theory is to count as one of them—offered a satisfactory basis for ethical decision making.

Natural law ethics. As far as secular moral philosophy is concerned, during most of the 20th century, natural law ethics has been considered a lifeless medieval relic, preserved only in Roman Catholic schools of moral theology. It is still true that the chief proponents of natural law are of that particular religious persuasion, but they have recently begun to defend their position by arguments that make no explicit appeal to their religious beliefs. Instead, they start their ethics with the claim that there are certain basic human goods that we should not act against. In the list offered by John Finnis in *Natural Law and Natural Rights* (1980), for example, these goods are life, knowledge, play, aesthetic experience, friendship, practical reasonableness, and religion. The identification of these goods is a matter of reflection, assisted by the findings of anthropologists. Each of the basic goods is regarded as equally fundamental; there is no hierarchy among them.

It would, of course, be possible to hold a consequentialist ethic that identified several basic human goods of equal importance and judged actions by their tendency to produce or maintain these goods. Thus, if life is a good, any action that led to a preventable loss of life would, other things being equal, be wrong. Natural law ethics, however, rejects this consequentialist approach. It makes the claim that it is impossible to measure the basic goods against each other. Instead of engaging in consequentialist calculations, the natural law ethic is built on the absolute prohibition of any action that aims directly against any basic good. The killing of the innocent, for instance, is always wrong, even if somehow killing one innocent person were to be the only way of saving thousands of innocent people. What is not adequately explained in this rejection of consequentialism is why the life of one innocent person—about whom, let us say, we know no more than that he is innocent—cannot be measured against the lives of a thousand innocent people about whom we have precisely the same information.

The basis of modern natural law ethic

Natural law ethics does allow one means of softening the effect of its absolute prohibitions. This is the doctrine of double effect, traditionally applied by Roman Catholic writers to some cases of abortion. If a pregnant woman is found to have a cancerous uterus, the doctrine of double effect allows a doctor to remove the uterus notwithstanding the fact that such action will kill the fetus. This allowance is made not because the life of the mother is regarded as more valuable than the life of the fetus, but because in removing the uterus the doctor is held not to aim directly at the death of the fetus. Instead, its death is an unwanted and indirect side effect of the laudable act of removing a diseased organ. On the other hand, a different medical condition might mean that the only way of saving the mother's life is by directly killing the fetus. Some years ago before the development of modern obstetric techniques, this was the case if the head of the fetus became lodged during delivery. Then the only way of saving the life of the woman was to crush the skull of the fetus. Such a procedure was prohibited, for in performing it the doctor would be directly killing the fetus. This ruling was applied even to those cases in which the death of the mother would certainly bring about the death of the fetus as well. The claim was that the doctor who killed the fetus directly was responsible for a murder, but the deaths from natural causes of the mother and fetus were not considered to be the doctor's doing. The example is significant because it indicates the lengths to which proponents of the natural law ethics are prepared to go in order to preserve the absolute nature of the prohibitions.

Ethical egoism. All of the normative theories considered so far have had a universal focus—*i.e.*, if they have been consequentialist theories, the goods they sought to achieve were sought for all capable of benefitting from them; and if they were deontological theories, the deontological principles applied equally to whoever might do the act in question. Ethical egoism departs from this consensus, suggesting that we should each consider only

Avoidance of conflict between morality and self-interest

the consequences of our actions for our own interests. The great advantage of such a position is that it avoids any possible conflict between morality and self-interest. If it is rational for us to pursue our own interest, then, if the ethical egoist is right, the rationality of morality is equally clear.

We can distinguish two forms of egoism. The individual egoist says, "Everyone should do what is in my interests." This indeed is egoism, but it is incapable of being couched in a universalizable form, and so it is arguably not a form of ethical egoism. Nor is the individual egoist likely to be able to persuade others to follow a course of action that is so obviously designed to benefit only the person who is advocating it.

Universal egoism is based on the principle "Everyone should do what is in her or his own interests." This principle is universalizable, since it contains no reference to any particular individual and it is clearly an ethical principle. Others may be disposed to accept it because it appears to offer them the surest possible way of furthering their own interests. Accordingly, this form of egoism is from time to time seized upon by some popular writer who proclaims it the obvious answer to all our ills and has no difficulty finding agreement from a segment of the general public. The U.S. writer Ayn Rand is perhaps the best 20th-century example. Rand's version of egoism is expounded in the novel *Atlas Shrugged* (1957) by her hero, John Galt, and in *The Virtue of Selfishness* (1965), a collection of her essays. It is a confusing mixture of appeals to self-interest and suggestions that everyone will benefit from the liberation of the creative energy that will flow from unfettered self-interest. Overlaying all this is the idea that true self-interest cannot be served by stealing, cheating, or similarly antisocial conduct.

As this example illustrates, what starts out as a defense of ethical egoism very often turns into an indirect form of Utilitarianism; the claim is that we will all be better off if each of us does what is in his or her own interest. The ethical egoist is virtually compelled to make this claim because otherwise there is a paradox in the fact that the ethical egoist advocates ethical egoism at all. Such advocacy would be contrary to the very principle of ethical egoism, unless the egoist benefits from others' becoming ethical egoists. If we see our interests as threatened by others' pursuing their own interests, we will certainly not benefit by others' becoming egoists; we would do better to keep our own belief in egoism secret and advocate altruism.

Unfortunately for ethical egoism, the claim that we will all be better off if every one of us does what is in his or her own interest is incorrect. This is shown by what are known as "prisoner's dilemma" situations, which are playing an increasingly important role in discussions of ethical theory. The basic prisoner's dilemma is an imaginary situation in which two prisoners are accused of a crime. If one confesses and the other does not, the prisoner who confesses will be released immediately and the other who does not will spend the next 20 years in prison. If neither confesses, each will be held for a few months and then both will be released. And if both confess, they will each be jailed for 15 years. The prisoners cannot communicate with one another. If each of them does a purely self-interested calculation, the result will be that it is better to confess than not to confess no matter what the other prisoner does. Paradoxical as it might seem, two prisoners, each pursuing his own interest, will end up worse than they would if they were not egoists.

The example might seem bizarre, but analogous situations occur quite frequently on a larger scale. Consider the dilemma of the commuter. Suppose that each commuter finds his or her private car a little more convenient than the bus; but when each of them drives a car, the traffic becomes so congested that everyone would be better off if they all took the bus and the buses moved quickly without traffic holdups. Because private cars are somewhat more convenient than buses, however, and the overall volume of traffic is not appreciably affected by one more car on the road, it is in the interest of each to continue using a private car. At least on the collective level, therefore, egoism is self-defeating—a conclusion well brought out by Parfit in his aforementioned *Reasons and Persons.*

APPLIED ETHICS

The most striking development in the study of ethics since the mid-1960s has been the growth of interest among philosophers in practical, or applied, ethics; *i.e.,* the application of normative theories to practical moral problems. This is not, admittedly, a totally new departure. From Plato onward moral philosophers have concerned themselves with practical questions, including suicide, the exposure of infants, the treatment of women, and the proper behaviour of public officials. Christian philosophers, notably Augustine and Aquinas, examined with great care such matters as when a war was just, whether it could ever be right to tell a lie, or if a Christian woman did wrong to commit suicide in order to save herself from rape. Hobbes had an eminently practical purpose in writing his *Leviathan,* and Hume wrote about the ethics of suicide. Practical concerns continued with the British Utilitarians, who saw reform as the aim of their philosophy: Bentham wrote on an incredible variety of topics, and Mill is celebrated for his essays on liberty and on the subjection of women.

The application of normative theories to practical moral problems

Nevertheless, during the first six decades of the 20th century moral philosophers largely isolated themselves from practical ethics—something that now seems all but incredible, considering the traumatic events through which most of them lived. There were one or two notable exceptions. The philosopher Bertrand Russell was very much involved in practical issues, but his stature among his colleagues was based on his work in logic and metaphysics and had nothing to do with his writings on topics such as disarmament and sexual morality. Russell himself seems to have regarded his practical contributions as largely separate from his philosophical work and did not develop his ethical views in any systematic or rigorous fashion.

The prevailing view of the period was that moral philosophy is quite separate from "moralizing," a task best left to preachers. What was not generally considered was whether moral philosophers could, without merely preaching, make an effective contribution to discussions of practical issues involving difficult ethical questions. The value of such work began to be widely recognized only during the 1960s, when first the U.S. civil rights movement and subsequently the Vietnam War and the rise of student activism started to draw philosophers into discussions of the moral issues of equality, justice, war, and civil disobedience. (Interestingly, there has been very little discussion of sexual morality—an indication that a subject once almost synonymous with the term morals has become marginal to our moral concerns.)

The founding, in 1971, of *Philosophy and Public Affairs,* a new journal devoted to the application of philosophy to public issues, provided both a forum and a new standard of rigour for these contributions. Applied ethics soon became part of the teaching of most philosophy departments of universities in English-speaking countries. Here it is not possible to do more than briefly mention some of the major areas of applied ethics and point to the issues that they raise.

Applications of equality. Since much of the early impetus for applied ethics came from the U.S. civil rights movement, such topics as equality, human rights, and justice have been prominent. We often make statements such as "All humans are equal" without thinking too deeply about the justification for the claims. Since the mid-1960s much has been written about how they can be justified. Discussions of this sort have led in several directions, often following social and political movements. The initial focus, especially in the United States, was on racial equality, and here, for once, there was a general consensus among philosophers on the unacceptability of discrimination against blacks. With so little disagreement about racial discrimination itself, the centre of attention soon moved to reverse discrimination: Is it acceptable to favour blacks for jobs and enrollment in universities and colleges because they had been discriminated against in the past and were generally so much worse off than

Questions related to racial discrimination

whites? Or is this, too, a form of racial discrimination and unacceptable for that reason?

Inequality between the sexes has been another focus of discussion. Does equality here mean ending as far as possible all differences in the sex roles, or could we have equal status for different roles? There has been a lively debate—both between feminists and their opponents and, on a different level, among feminists themselves—about what a society without sexual inequality would be like. Here, too, the legitimacy of reverse discrimination has been a contentious issue. Feminist philosophers have also been involved in debates over abortion and new methods of reproduction. These topics will be covered separately below.

Many discussions of justice and equality are limited in scope to a single society. Even Rawls's theory of justice, for example, has nothing to say about the distribution of wealth between societies, a subject that could make acceptance of his maximin principle much more onerous. But philosophers have now begun to think about the moral implications of the inequality in wealth between the affluent nations (and their citizens) and those living in countries subject to famine. What are the obligations of those who have plenty when others are starving? It has not proved difficult to make a strong case for the view that affluent nations, as well as affluent individuals, ought to be doing much more to help the poor than they are generally now doing.

There is one issue related to equality in which philosophers have led, rather than followed, a social movement. In the early 1970s, a group of young Oxford-based philosophers began to question the assumption that while all humans are entitled to equal moral status, nonhuman animals automatically have an inferior position. The publication in 1972 of *Animals, Men and Morals: An Inquiry into the Maltreatment of Non-humans,* edited by Roslind and Stanley Godlovitch and John Harris, was followed three years later by Peter Singer's *Animal Liberation* and then by a flood of articles and books that established the issue as a part of applied ethics. At the same time, these writings provided the philosophical basis for the animal liberation movement, which has had an effect on attitudes and practices toward animals in many countries.

Environmental ethics. Environmental issues raise a host of difficult ethical questions, including the ancient one of the nature of intrinsic value. Whereas many philosophers in the past have agreed that human experiences have intrinsic value and the Utilitarians at least have always accepted that the pleasures and pains of nonhuman animals are of some intrinsic significance, this does not show why it is so bad if dodos become extinct or a rain forest is cut down. Are these things to be regretted only because of the loss to humans or other sentient creatures? Or is there more to it than that? Some philosophers are now prepared to defend the view that trees, rivers, species (considered apart from the individual animals of which they consist), and perhaps ecological systems as a whole have a value independent of the instrumental value they may have for humans or other sentient creatures.

Issue of obligations to future generations

Our concern for the environment also raises the question of our obligations to future generations. How much do we owe to the future? From a social contract view of ethics or for the ethical egoist, the answer would seem to be: nothing. For we can benefit them, but they are unable to reciprocate. Most other ethical theories, however, do give weight to the interests of coming generations. Utilitarians, for one, would not think that the fact that members of future generations do not exist yet is any reason for giving less consideration to their interests than we give to our own, provided only that we are certain that they will exist and will have interests that will be affected by what we do. In the case of, say, the storage of radioactive wastes, it seems clear that what we do will indeed affect the interests of generations to come.

The question becomes much more complex, however, when we consider that we can affect the size of future generations by the population policies we choose and the extent to which we encourage large or small families. Most environmentalists believe that the world is already dangerously overcrowded. This may well be so, but the notion of overpopulation conceals a philosophical issue that is ingeniously explored by Derek Parfit in *Reasons and Persons* (1984). What is optimum population? Is it that population size at which the average level of welfare will be as high as possible? Or is it the size at which the total amount of welfare—the average multiplied by the number of people—is as great as possible? Both answers lead to counterintuitive outcomes, and the question remains one of the most baffling mysteries in applied ethics.

War and peace. The Vietnam War ensured that discussions as to the justness of a war and of the legitimacy of conscription and civil disobedience were prominent in early writings in applied ethics. There was considerable support for civil disobedience against unjust aggression and against unjust laws even in a democracy.

With the cessation of hostilities in Vietnam and the end of conscription, interest in these questions declined. Concern about nuclear weapons in the early 1980s, however, has caused philosophers to argue about whether nuclear deterrence can be an ethically acceptable strategy if it means using civilian populations as potential nuclear targets. Jonathan Schell's *Fate of the Earth* (1982) raised several philosophical questions about what we ought to do in the face of the possible destruction of all life on our planet.

Abortion, euthanasia, and the value of human life. A number of ethical questions cluster around both ends of the human life span. Whether abortion is morally justifiable has popularly been seen as depending on our answer to the question "When does a human life begin?" Many philosophers believe this to be the wrong question to ask because it suggests that there might be a factual answer that we can somehow discover through advances in science. Instead, these philosophers think we need to ask what it is that makes killing a human being wrong and then consider whether these characteristics, whatever they might be, apply to the fetus in an abortion. There is no generally agreed upon answer, yet some philosophers have presented surprisingly strong arguments to the effect that not only the fetus but even the newborn infant has no right to life. This position has been defended by Jonathan Glover in *Causing Death and Saving Lives* (1977) and in more detail by Michael Tooley in *Abortion and Infanticide* (1984).

Such views have been hotly contested, especially by those who claim that all human life, irrespective of its characteristics, must be regarded as sacrosanct. The task for those who defend the sanctity of human life is to explain why human life, no matter what its characteristics, is specially worthy of protection. Explanation could no doubt be provided in terms of such traditional Christian doctrines as that all humans are made in the image of God or that all humans have an immortal soul. In the current debate, however, the opponents of abortion have eschewed religious arguments of this kind without finding a convincing secular alternative.

Somewhat similar issues are raised by euthanasia when it is nonvoluntary, as, for example, in the case of severely disabled newborn infants. Euthanasia, however, can be voluntary, and this has brought it support from some who hold that the state should not interfere with the free, informed choices of its citizens in matters that do not cause others harm. (The same argument is often invoked in defense of the pro-choice position in the abortion controversy; but it is on much weaker ground in this case because it presupposes what it needs to prove—namely, that the fetus does not count as an "other.") Opposition to voluntary euthanasia has centred on practical matters such as the difficulty of adequate safeguards and on the argument that it would lead to a "slippery slope" that would take us to nonvoluntary euthanasia and eventually to the compulsory involuntary killing of those the state considers to be socially undesirable.

Distinction between killing and allowing to die

Philosophers have also canvassed the moral significance of the distinction between killing and allowing to die, which is reflected in the fact that many physicians will allow a patient with an incurable condition to die when life could still be prolonged, but they will not take active steps to end the patient's life. Consequentialist philosophers, among them both Glover and Tooley, have denied that

this distinction possesses any intrinsic moral significance. For those who uphold a system of absolute rules, on the other hand, a distinction between acts and omissions is essential if they are to render plausible the claim that we must never breach a valid moral rule.

Bioethics. The issues of abortion and euthanasia are included in one of the fastest growing areas of applied ethics, that dealing with ethical issues raised by new developments in medicine and the biological sciences. This subject, known as bioethics, often involves interdisciplinary work, with physicians, lawyers, scientists, and theologians all taking part. Centres for research in bioethics have been established in Australia, Britain, Canada, and the United States. Many medical schools have added the discussion of ethical issues in medicine to their curricula. Governments have sought to deal with the most controversial issues by appointing special committees to provide ethical advice.

Major issues of bioethics

Several key themes run through the subjects covered by bioethics. One, related to abortion and euthanasia, is whether the quality of a human life can be a reason for ending it or for deciding not to take steps to prolong it. Since medical science can now keep alive severely disabled infants who a few years ago would have died soon after birth, pediatricians are regularly faced with this question. The issue received national publicity in Britain in 1981 when a respected pediatrician was charged with murder, following the death of an infant with Down's syndrome. Evidence at the trial indicated that the parents had not wanted the child to live and that the pediatrician had consequently prescribed a narcotic painkiller. The doctor was acquitted. The following year, in the United States, an even greater furor was caused by a doctor's decision to follow the wishes of the parents of a Down's syndrome infant and not carry out surgery without which the baby would die. The doctor's decision was upheld by the Supreme Court of Indiana, and the baby died before an appeal could be made to the U.S. Supreme Court. In spite of the controversy and efforts by government officials to ensure that handicapped infants are given all necessary lifesaving treatment, in neither Britain nor the United States is there any consensus about the decisions that should be made when severely disabled infants are born or by whom these decisions should be made.

Medical advances have raised other related questions. Even those who defend the doctrine of the sanctity of all human life do not believe that doctors have to use extraordinary means to prolong life, but the distinction between ordinary and extraordinary means, like that between acts and omissions, is itself under attack. Critics assert that the wishes of the patient or, if these cannot be ascertained, the quality of the patient's life provides a more relevant basis for a decision than the nature of the means to be used.

Another central theme is that of patient autonomy. This arises not only in the case of voluntary euthanasia but also in the area of human experimentation, which has come under close scrutiny following reported abuses. It is generally agreed that patients must give informed consent to any experimental procedures. But how much and how detailed information is the patient to be given? The problem is particularly acute in the case of randomly controlled trials, which scientists consider the most desirable way of testing the efficacy of a new procedure but which require that the patient agree to being administered randomly one of two or more forms of treatment.

The allocation of medical resources became a life-and-death issue when hospitals obtained dialysis machines and had to choose which of their patients suffering from kidney disease would be able to use the scarce machines. Some argued for "first come, first served," whereas others thought it obvious that younger patients or patients with dependents should have preference. Kidney machines are no longer as scarce, but the availability of various other exotic, expensive lifesaving techniques is limited; hence, the search for rational principles of distribution continues.

New issues arise as further advances are made in biology and medicine. In 1978 the birth of the first human being to be conceived outside the human body initiated a debate about the ethics of in vitro fertilization. This soon led to questions about the freezing of human embryos and what should be done with them if, as happened in 1984 with two embryos frozen by an Australian medical team, the parents should die. The next controversy in this area arose over commercial agencies offering infertile married couples a surrogate mother who would for a fee be impregnated with the sperm of the husband and then surrender the resulting baby to the couple. Several questions emerged: Should we allow women to rent their wombs to the highest bidder? If a woman who has agreed to act as a surrogate changes her mind and decides to keep the baby, should she be allowed to do so?

The culmination of such advances in human reproduction will be the mastery of genetic engineering. Then we will all face the question posed by the title of Jonathan Glover's probing book *What Sort of People Should There Be?* (1984). Perhaps this will be the most challenging issue for 21st-century ethics.

BIBLIOGRAPHY

General works: For an introduction to the major theories of ethics, the reader should consult RICHARD B. BRANDT, *Ethical Theory: The Problems of Normative and Critical Ethics* (1959), an excellent comprehensive textbook. WILLIAM K. FRANKENA, *Ethics,* 2nd ed. (1973), is a much briefer treatment. Another concise work is BERNARD WILLIAMS, *Ethics and the Limits of Philosophy* (1985). There are several useful collections of classical and modern writings; among the better ones are OLIVER A. JOHNSON, *Ethics: Selections from Classical and Contemporary Writers,* 5th ed. (1984); and JAMES RACHELS (ed.), *Understanding Moral Philosophy* (1976), which places greater emphasis on modern writers.

Origins of ethics: JOYCE O. HERTZLER, *The Social Thought of the Ancient Civilizations* (1936, reissued 1961), is a wide-ranging collection of materials. EDWARD WESTERMARCK, *The Origin and Development of the Moral Ideas,* 2 vol., 2nd ed. (1912–17, reprinted 1971), is dated but still unsurpassed as a comprehensive account of anthropological data. MARY MIDGLEY, *Beast and Man: The Roots of Human Nature* (1978, reissued 1980), is excellent on the links between biology and ethics; and EDWARD O. WILSON, *Sociobiology: The New Synthesis* (1975), and *On Human Nature* (1978), contain controversial speculations on the biological basis of social behaviour. RICHARD DAWKINS, *The Selfish Gene* (1976, reprinted 1978), is another evolutionary account, fascinating but to be used with care.

History of Western ethics: HENRY SIDGWICK, *Outlines of the History of Ethics for English Readers,* 6th enlarged ed. (1931, reissued 1967), is a triumph of scholarship and brevity. WILLIAM EDWARD HARTPOLE LECKY, *History of European Morals from Augustus to Charlemagne,* 2 vol., 3rd rev. ed. (1877, reprinted 1975), is fascinating and informative. Among more recent histories, VERNON J. BOURKE, *History of Ethics* (1968, reissued in 2 vol., 1970), is remarkably comprehensive; while ALASDAIRE MACINTYRE, *A Short History of Ethics* (1966), is a readable personal view.

Indian ethics: SURAMA DASGUPTA, *Development of Moral Philosophy in India* (1961, reissued 1965), is a clear discussion of the various schools. SARVEPALLI RADHAKRISHNAN and CHARLES A. MOORE (eds.), *A Source Book in Indian Philosophy* (1957, reprinted 1967), is a collection of key primary sources. For Buddhist texts, see EDWARD CONZE *et al.* (eds.), *Buddhist Texts Through the Ages* (1954, reissued 1964).

Chinese ethics: Standard introductions to the works of classic Chinese authors mentioned in the article are E.R. HUGHES (ed.), *Chinese Philosophy in Classical Times* (1942, reprinted 1966); and FUNG YU-LAN, *A History of Chinese Philosophy,* 2 vol., trans. from the Chinese (1952–53, reprinted 1983).

Ancient Greek and Roman ethics: JONATHAN BARNES, *The Presocratic Philosophers,* rev. ed. (1982), treats Greek ethics before Socrates. The central texts of the Classic period of Greek ethics are PLATO, *Politeia* (*The Republic*), *Euthyphro, Protagoras,* and *Gorgias;* and ARISTOTLE, *Ethica Nicomachea* (*Nicomachean Ethics*). A concise introduction to the ethical thought of this period is provided by PAMELA HUBY, *Greek Ethics* (1967); and CHRISTOPHER ROWE, *An Introduction to Greek Ethics* (1976). Significant writings of the Stoics include MARCUS TULLIUS CICERO, *De officiis* (*On Duties*); LUCIUS ANNAEUS SENECA, *Epistulae morales* (*Moral Essays*); and MARCUS AURELIUS, *D. imperatoris Marci Antonini Commentariorum qvos sibi ipsi scripsit libri XII* (*The Meditations of the Emperor Marcus Antoninus*). From Epicurus only fragments remain; they have been collected in CYRIL BAILEY (ed.), *Epicurus, the Extant Remains* (1926, reprinted 1979). The most complete of the surviving works of the Epicureans is LUCRETIUS, *De rerum natura* (*On the Nature of Things*).

Early and medieval Christian ethics: In addition to the

Gospels and Paul's letters, important writings include ST. AUGUSTINE, *De civitate Dei* (413–426; *The City of God*), and *Enchiridion ad Laurentium de fide, spe, et caritate* (421; *Enchiridion to Laurentius on Faith, Hope and Love*); PETER ABELARD, *Ethica* (*c.* 1135; *Ethics*); and ST. THOMAS AQUINAS, *Summa theologiae* (1265 or 1266–73). On the history of the transition from Roman ethics to Christianity, W.E.H. LECKY, *op.cit.*, remains unsurpassed. D.J. O'CONNOR, *Aquinas and Natural Law* (1967), is a brief introduction to the most important of the Scholastic writers on ethics.

Ethics of the Renaissance and Reformation: Machiavelli's chief works are available in modern translations: NICCOLÒ MACHIAVELLI, *The Prince,* trans. and ed. by PETER BONDANELLA and MARK MUSA (1984), and *The Discourses,* trans. by LESLIE J. WALKER (1975). For Luther's writings, see the comprehensive edition MARTIN LUTHER, *Works,* 55 vol., ed. by JAROSLAV PELIKAN *et al.* (1955–76). Calvin's major work is available in JEAN CALVIN, *Institutes of the Christian Religion,* trans. by HENRY BEVERIDGE, 2 vol. (1979).

The British tradition from Hobbes to the Utilitarians: The key works of this period include THOMAS HOBBES, *Leviathan* (1651); RALPH CUDWORTH, *Eternal and Immutable Morality* (published posthumously, 1688); HENRY MORE, *Enchiridion Ethicum* (1662); SAMUEL CLARKE, Boyle lectures for 1705, published in his *Works,* 4 vol. (1738–42); 3RD EARL OF SHAFTESBURY, "Inquiry Concerning Virtue or Merit," published together with other essays in his *Characteristicks of Men, Manners, Opinions, Times* (1711); JOSEPH BUTLER, *Fifteen Sermons* (1726); FRANCIS HUTCHESON, *Inquiry into the Original of Our Ideas of Beauty and Virtue* (1725), and *A System of Moral Philosophy,* 2 vol. (1755); DAVID HUME, *A Treatise of Human Nature* (1739–40), and *An Enquiry Concerning the Principles of Morals* (1751); RICHARD PRICE, *A Review of the Principal Questions and Difficulties in Morals* (1758); THOMAS REID, *Essays on the Active Powers of the Human Mind* (1758); WILLIAM PALEY, *The Principles of Moral and Political Philosophy* (1785); JEREMY BENTHAM, *Introduction to the Principles of Morals and Legislation* (1789); JOHN STUART MILL, *Utilitarianism* (1863); and HENRY SIDGWICK, *The Methods of Ethics* (1874). Selections of the major texts of this period are brought together in D.D. RAPHAEL (ed.), *British Moralists, 1650–1800,* 2 vol. (1969); and in D.H. MONRO (ed.), *A Guide to the British Moralists* (1972). Useful introductions to separate writers include J. KEMP, *Ethical Naturalism* (1970), on Hobbes and Hume; W.D. HUDSON, *Ethical Intuitionism* (1967), on the intuitionists from Cudworth to Price and the debate with the moral sense school; and ANTHONY QUINTON, *Utilitarian Ethics* (1973). C.D. BROAD, *Five Types of Ethical Theory* (1930, reprinted 1971), includes clear accounts of the ethics of Butler, Hume, and Sidgwick. J.L. MACKIE, *Hume's Moral Theory* (1980), brilliantly traces the relevance of Hume's work to current disputes about the nature of ethics.

The continental tradition from Spinoza to Nietzsche: The major texts are available in many English translations. See BARUCH SPINOZA, *The Ethics and Selected Letters,* trans. by SAMUEL SHIRLEY, ed. by SEYMOUR FELDMAN (1982); JEAN-JACQUES ROUSSEAU, *A Discourse on Inequality,* trans. by MAURICE CRANSTON (1984), and *The Social Contract,* annotated ed., trans. by CHARLES M. SHEROVER (1974); IMMANUEL KANT, *Grounding for the Metaphysics of Morals,* trans. by JAMES W. ELLINGTON (1981), and *Critique of Practical Reason, and Other Writings in Moral Philosophy,* ed. and trans. by LEWIS WHITE BECK (1949, reprinted 1976); G.W.F. HEGEL, *Phenomenology of Spirit,* trans. by A.V. MILLER (1977), and *Hegel's Philosophy of Right,* trans. by T.M. KNOX (1967, reprinted 1980); KARL MARX, *Economic and Philosophic Manuscripts of 1844,* ed. by DIRK J. STRUIK (1964), *Capital: A Critique of Political Economy,* trans. by DAVID FERNBACH, 3 vol. (1981), and *The Communist Manifesto of Marx and Engels,* ed. by HAROLD J. LASKI (1967, reprinted 1975); FRIEDRICH NIETZSCHE, *Beyond Good and Evil: Prelude to a Philosophy of the Future,* trans. by R.J. HOLLINGDALE (1973), and *The Genealogy of Morals: A Polemic,* trans. by HORACE B. SAMUEL (1964). Among the easier introductory studies are H.B. ACTON, *Kant's Moral Philosophy* (1970); and PETER SINGER, *Hegel* (1983), and *Marx* (1980). C.D. BROAD, *op. cit.*, contains readable accounts of the ethics of both Spinoza and Kant.

20th-century Western ethics: The most influential writings in metaethics during the 20th century have been GEORGE EDWARD MOORE, *Principia Ethica* (1903, reprinted 1976); W.D. ROSS, *The Right and the Good* (1930, reprinted 1973); A.J. AYER, *Language, Truth, and Logic* (1936, reissued 1974); CHARLES L. STEVENSON, *Ethics and Language* (1944, reprinted 1979); R.M. HARE, *The Language of Morals* (1952, reprinted 1972), and *Freedom and Reason* (1963, reprinted 1977); and, in France, JEAN-PAUL SARTRE, *Being and Nothingness* (1956, reissued 1978; originally published in French, 1943), and *Existentialism and Humanism* (1948, reprinted 1977; originally published in French, 1946). RALPH BARTON PERRY, *General Theory of Value* (1926, reprinted 1967), was highly regarded in the United States but comparatively neglected elsewhere. WILFRID SELLARS and JOHN HOSPERS (eds.), *Readings in Ethical History,* 2nd ed. (1970), contains the most important pieces of writing on ethics from the first half of the 20th century. Widely discussed later works include THOMAS NAGEL, *The Possibility of Altruism* (1970, reissued 1978); G.J. WARNOCK, *The Object of Morality* (1971); J.L. MACKIE, *Ethics: Inventing Right and Wrong* (1977); RICHARD B. BRANDT, *A Theory of the Good and the Right* (1979); JOHN FINNIS, *Natural Law and Natural Rights* (1980); and R.M. HARE, *Moral Thinking: Its Levels, Method, and Point* (1981). A defense of naturalism can be found in two important articles by PHILIPPA FOOT, "Moral Beliefs" and "Moral Arguments," both originally published in 1958 and later reprinted in her *Virtues and Vices, and Other Essays in Moral Philosophy* (1978, reprinted 1981). DAVID WIGGINS, *Truth, Invention, and the Meaning of Life* (1976), is a statement of what has come to be known as "moral realism." MARY WARNOCK, *Ethics Since 1900,* 3rd ed. (1978); G.J. WARNOCK, *Contemporary Moral Philosophy* (1967); and W.D. HUDSON, *A Century of Moral Philosophy* (1980), provide guidance through 20th-century metaethical disputes.

Normative ethics: For Moore's ideal Utilitarianism, see G.E. MOORE, *Ethics,* 2nd ed. (1966). The best short statement of an act-Utilitarian position is J.J.C. Smart's contribution to J.J.C. SMART and BERNARD WILLIAMS, *Utilitarianism: For and Against* (1973). R.M. HARE, *op. cit.,* is an extended argument for a form of preference Utilitarianism that allows some scope to moral principles while not departing from act-Utilitarianism at the level of critical thought. DAVID LYONS, *Forms and Limits of Utilitarianism* (1965), probes the distinction between act- and rule-Utilitarianism. RICHARD B. BRANDT, *op. cit.,* includes a defense of a version of rule-Utilitarianism. DONALD REGAN, *Utilitarianism and Co-operation* (1980), is an ingenious discussion of how the need to cooperate can be incorporated into Utilitarian theory. AMARTYA SEN and BERNARD WILLIAMS (eds.), *Utilitarianism and Beyond* (1982), is a collection of essays on the difficulties of the Utilitarian position. A major contribution to consequentialist theory is DEREK PARFIT, *Reasons and Persons* (1984), which includes penetrating arguments on the nature of consequentialist reasoning in ethics. The standard defense of an ethic of prima facie duties remains W.D. ROSS, *op. cit.* H.J. MCCLOSKEY, *Meta-Ethics and Normative Ethics* (1969), is a restatement with some modifications. The most widely discussed alternative theory to Utilitarianism in recent years is set forth in JOHN RAWLS, *A Theory of Justice* (1971, reprinted 1981). ROBERT NOZICK, *Anarchy, State, and Utopia* (1974), criticizes Rawls and presents a rights-based theory. Another work giving prominence to rights is RONALD DWORKIN, *Taking Rights Seriously* (1977). Very different from the approach of both Nozick and Dworkin is the attempt to ground rights in natural law in JOHN FINNIS, *op. cit.,* and a shorter and more accessible introduction to natural law ethics is *Fundamentals of Ethics* (1983). Egoism as a theory of rationality is discussed by DEREK PARFIT, *op. cit.;* a useful collection of readings on this topic is DAVID P. GAUTHIER (ed.), *Morality and Rational Self-Interest* (1970); see also RONALD D. MILO (ed.), *Egoism and Altruism* (1973).

Applied ethics: Many of the best examples of applied ethics are to be found in journal articles, particularly in *Philosophy and Public Affairs* (quarterly). There are many anthologies of representative samples of such writings. Among the better ones are JAMES RACHELS (ed.), *Moral Problems,* 3rd ed. (1979); JAN NARVESON (ed.), *Moral Issues* (1983); and MANUEL VELASQUEZ and CYNTHIA ROSTANKOWSKI, *Ethics, Theory and Practice* (1985). There are also books and collections on specific topics. MARSHALL COHEN, THOMAS NAGEL, and THOMAS SCANLON (eds.), *Equality and Preferential Treatment* (1977), is a collection of some of the best articles on equality and reverse discrimination; while ALAN H. GOLDMAN, *Justice and Reverse Discrimination* (1979), is a book-length treatment of the issues. Some of the more philosophically probing discussions of feminism are JANET RADCLIFFE RICHARDS, *The Sceptical Feminist* (1980, reprinted with corrections, 1982); MARY MIDGLEY and JUDITH HUGHES, *Women's Choices: Philosophical Problems Facing Feminism* (1983); and ALISON M. JAGGAR, *Feminist Politics and Human Nature* (1983). The moral obligations of the wealthy toward the starving are discussed in the anthology *World Hunger and Moral Obligation,* ed. by WILLIAM AIKEN and HUGH LAFOLLETTE.

The ethics of the treatment of animals has given rise to much philosophical discussion. Books arguing for radical change include STANLEY GODLOVITCH, ROSLIND GODLOVITCH, and JOHN HARRIS (eds.), *Animals, Man, and Morals: An Enquiry into the Maltreatment of Non-Humans* (1971); PETER SINGER, *Animal Liberation: A New Ethics for Our Treatment of Animals* (1975); STEPHEN R.L. CLARK, *The Moral Status of Animals* (1977, reis-

sued 1984); and TOM REGAN, *The Case for Animal Rights* (1983). R.G. FREY, *Interests and Rights: The Case Against Animals* (1980), and *Rights, Killing, and Suffering: Moral Vegetarianism and Applied Ethics* (1983), resist some of these arguments. MARY MIDGLEY, *Animals and Why They Matter* (1983), takes a middle course.

Essays dealing with ethical issues raised by concern for the environment are collected in ROBERT ELLIOT and ARRAN GARE (eds.), *Environmental Philosophy* (1983); and K.S. SHRADER-FRECHETTE, *Environmental Ethics* (1981). Useful full-length studies include JOHN PASSMORE, *Man's Responsibility for Nature: Ecological Problems and Western Tradition,* 2nd ed. (1980); and H.J. MCCLOSKEY, *Ecological Ethics and Politics* (1983). For specific problems of future generations, see R. SIKORA and BRIAN BARRY (eds.), *Obligations to Future Generations* (1979). A difficult but fascinating discussion of the problem of optimum population size in an ideal world can be found in DEREK PARFIT, *op. cit.*

MICHAEL WALZER, *Just and Unjust Wars* (1977), is a fine study of the morality of war; RICHARD A. WASSERSTROM (ed.), *War and Morality* (1970), is a valuable collection of essays. NIGEL BLAKE and KAY POLE (eds.), *Objections to Nuclear Defence* (1984), and *Dangers of Deterrence* (1984), are collections of philosophical writings on nuclear war.

There is an immense amount of literature on abortion, though of various philosophical depth. MICHAEL TOOLEY, *Abortion and Infanticide* (1983), is a penetrating study. For contrasting views, see GERMAIN G. GRISEZ, *Abortion: The Myths, the Realities, and the Arguments* (1970); and BARUCH A. BRODY, *Abortion and the Sanctity of Human Life: A Philosophical View* (1975). Another notable treatment is L.W. SUMNER, *Abortion and Moral Theory* (1981). JOEL FEINBERG (ed.), *The Problem of Abortion,* 2nd ed. (1984), is a good collection of essays. For a discussion of sanctity of life issues in general, including both abortion and euthanasia, see JONATHAN GLOVER, *Causing Death and Saving Lives* (1977); and PETER SINGER, *Practical Ethics* (1979). The specific problem of the treatment of severely handicapped infants is discussed in HELGA KUHSE and PETER SINGER, *Should the Baby Live?* (1985).

For a comprehensive textbook on bioethics, see TOM. L. BEAUCHAMP and JAMES F. CHILDRESS, *Principles of Biomedical Ethics,* 2nd ed. (1983). Anthologies of essays on diverse topics in bioethics include SAMUEL GOROVITZ *et al.* (eds.), *Moral Problems in Medicine,* 2nd ed. (1983); and JOHN ARRAS and ROBERT HUNT (comp.), *Ethical Issues in Modern Medicine,* 2nd ed. (1983). JAMES F. CHILDRESS, *Who Should Decide?* (1982), deals with paternalism in medical care; while PETER SINGER and DEANE WELLS, *The Reproduction Revolution: New Ways of Making Babies* (1984), focusses on the new reproductive technology. For the philosophical issues underlying genetic engineering and other methods of altering the human organism, see JONATHAN GLOVER, *What Sort of People Should There Be?* (1984).

(P.Si.)

Europe

Among the continents, Europe is an anomaly. Larger only than Australia, it is a small appendage of the great landmass that it shares with an Asia more than four times its size. Yet the peninsular and insular western extremity of Eurasia, thrusting toward the North Atlantic Ocean, provides—thanks to its latitude and its physical geography—a relatively genial human habitat, and the long processes of human history came to mark off the region as the home of a distinctive civilization. In spite of its internal diversity, Europe has thus functioned, from the time it first emerged in the human consciousness, as a world apart, concentrating—to borrow a phrase from Christopher Marlowe—"infinite riches in a little room."

All the continents are conceptual constructs, but only Europe was not first perceived and named by outsiders. "Europa," as the more learned of the ancient Greeks first conceived it, stood in sharp contrast to both Asia and Libya, the name then applied to the known northern part of Africa. Literally, "Europa" is now thought to have meant "Mainland," rather than the earlier interpretation, "Sunset." It appears to have suggested itself to the Greeks, in their maritime world, as an appropriate designation of the broadening, extensive northerly lands that lay beyond, lands with characteristics but vaguely known; yet these characteristics were clearly different from those inherent in the concepts of Asia and Libya, both of which, relatively prosperous and civilized, were associated closely with the culture of the Greeks and their predecessors. Traders and travelers reported that Europe possessed distinctive physical units, with mountain systems and lowland river basins much larger than those familiar to inhabitants of the Mediterranean region. It was also clear that a succession of climates, markedly different from those of the Mediterranean borderlands, were to be experienced as Europe was penetrated from the south. The spacious eastern steppe and, to the west and north, primeval forests as yet only marginally touched by human occupancy further underlined environmental contrasts. Europe was culturally backward and scantily settled. It was a "barbarian"—that is, a non-Greek—world, its inhabitants making "bar-bar" noises in unintelligible tongues.

The Roman Empire, at its greatest extent in the 2nd century AD, revealed, and imprinted its culture on, much of the face of the continent, while trading relations beyond its frontiers also drew the remoter regions into its sphere. Yet it was not until the 19th and 20th centuries that modern science was able to draw with some precision the geologic and geographic lineaments of the European continent, the peoples of which had meanwhile achieved domination over—and set in motion vast countervailing movements among—the inhabitants of much of the rest of the globe.

As to the territorial limits of Europe, while these seem clear on its three seaward flanks, they have been uncertain and hence much debated on the east, where the continent merges, without sundering physical limits, with parts of western Asia. Even to the north and west, many island groups—Svalbard (Spitsbergen), the British Isles, the Faeroes, Iceland, and the Madeira and Canary islands—that are European by culture are included in the continent, although Greenland is conventionally allocated to North America. Further, the Mediterranean coastlands of North Africa and southwestern Asia also exhibit some European physical and cultural affinities, and Turkey and Cyprus, while geologically Asian, possess elements of European culture and may, perhaps, be regarded as parts of Europe. Eastward limits, now adopted by European (including Soviet) geographers, assign the Caucasus to Asia and are taken to run southward along the eastern foot of the Urals and then across the Mugodzhar Hills, along the Emba River, and along the northern shore of the Caspian Sea. West of the Caspian, the European limit follows the Kumo–Manych Depression and the Kerch Strait to the Black Sea.

This conventional eastern boundary, however, is not a cultural, political, or economic discontinuity on the land comparable, for example, to the insulating significance of the sparsely inhabited Himalayas, which clearly mark a northern limit to South Asian civilization. Inhabited plains, with only the minor interruption of the worn-down Ural Mountains, extend from central Europe to the Yenisey River in central Siberia. A homogeneous, highly centralized, Russian-based civilization dominates Soviet territory from the Baltic and Black seas to the Pacific Ocean. This civilization is distinguished from the rest of Europe by legacies of a medieval Mongol–Tatar domination that precluded sharing many of the innovations and developments of European "Western civilization." In partitioning the globe into meaningful large geographic units, therefore, modern geographers have been mindful of the distinctiveness of the Soviet Union. Most treat it as a distinct territorial entity, comparable to a continent, that is separate from Europe to the west and from the rest of Asia to the south and east. The ensuing discussion of Europe focuses primarily upon the territories and peoples lying west of the Soviet border, although note is taken, where appropriate, of physical and cultural features shared by the "European" Soviet Union with the rest of the continent. For a detailed discussion of the Soviet Union, see UNION OF SOVIET SOCIALIST REPUBLICS.

Europe occupies some four million square miles (10.4 million square kilometres) within the conventional borders assigned to it. Of this total, 2.1 million square miles are inside the Soviet Union. This broad territory reveals no simple unity of geologic structure, landform, relief, or climate. Rocks of all geologic periods are exposed, and the operation of geologic forces during an immense succession of eras has contributed to the molding of the landscapes of mountain, plateau, and lowland and has bequeathed a variety of mineral reserves. Glaciation, too, has left its mark over wide areas, and the processes of erosion and deposition have created a highly variegated and compartmentalized countryside. Climatically, Europe benefits by having only a small proportion of its surface either too cold or too hot and dry for effective settlement and use. Regional climatic contrasts nevertheless exist: oceanic, Mediterranean, and continental types occur widely, as do gradations from one to the other. Associated vegetation and soil forms also show continual variety, but little is left of the dominant woodland that clothed the continent when humans first appeared.

All in all, Europe enjoys a considerable and long-exploited resource base of soil, forest, sea, and minerals, notably coal, but population, considerable numerically, as well as technically highly qualified, is increasingly its principal resource. The continent contains a shrinking seventh portion of mankind, but this represents a population of high skill and initiative. Europe thus supports high densities of population, markedly concentrated in industrialized regions. In manufacture, commerce, and agriculture it still occupies an eminent, if no longer necessarily predominant, position, and, as agriculture increasingly rationalizes its structure, city life is, everywhere, becoming the norm.

Europe is preeminently the homeland of white peoples. Its early and continuing economic achievements, evidenced by a high standard of living, and its successes in science, technology, and the arts spring from the vigour of its peoples in developing a high civilization, the roots of which lie in ancient Greece and Rome, the Byzantine Empire, and Palestine. Whatever its indebtedness, Europe has always shown its own powers of creativity and leadership: although wracked and exhausted by continued

internal conflict, it has nevertheless advanced sufficiently to leave as its heritage the exploration, colonization, and development of other peoples and regions of the globe, if not always to the benefit of the other peoples and regions.

(W.G.E./T.M.P.)

This article treats the physical and human geography of Europe, followed by discussion of geographic features of special interest. For discussion of individual countries of the continent, see specific articles by name—*e.g.,* ITALY, POLAND, and UNITED KINGDOM. For discussion of major cities of the continent, see specific articles by name—*e.g.,* LONDON, ROME, and WARSAW. The principal articles discussing the historical and cultural development of the continent include EUROPEAN HISTORY AND CULTURE; GREEK AND ROMAN CIVILIZATIONS, ANCIENT; and HOLY ROMAN EMPIRE, THE HISTORY OF THE. Related topics are discussed in such articles as those on religion (*e.g.,* EUROPEAN RELIGIONS, ANCIENT; JUDAISM; and ROMAN CATHOLICISM) and literature (*e.g.,* DUTCH LITERATURE; HOMERIC EPICS, THE; and SPANISH LITERATURE). For further references, see the *Index.*

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

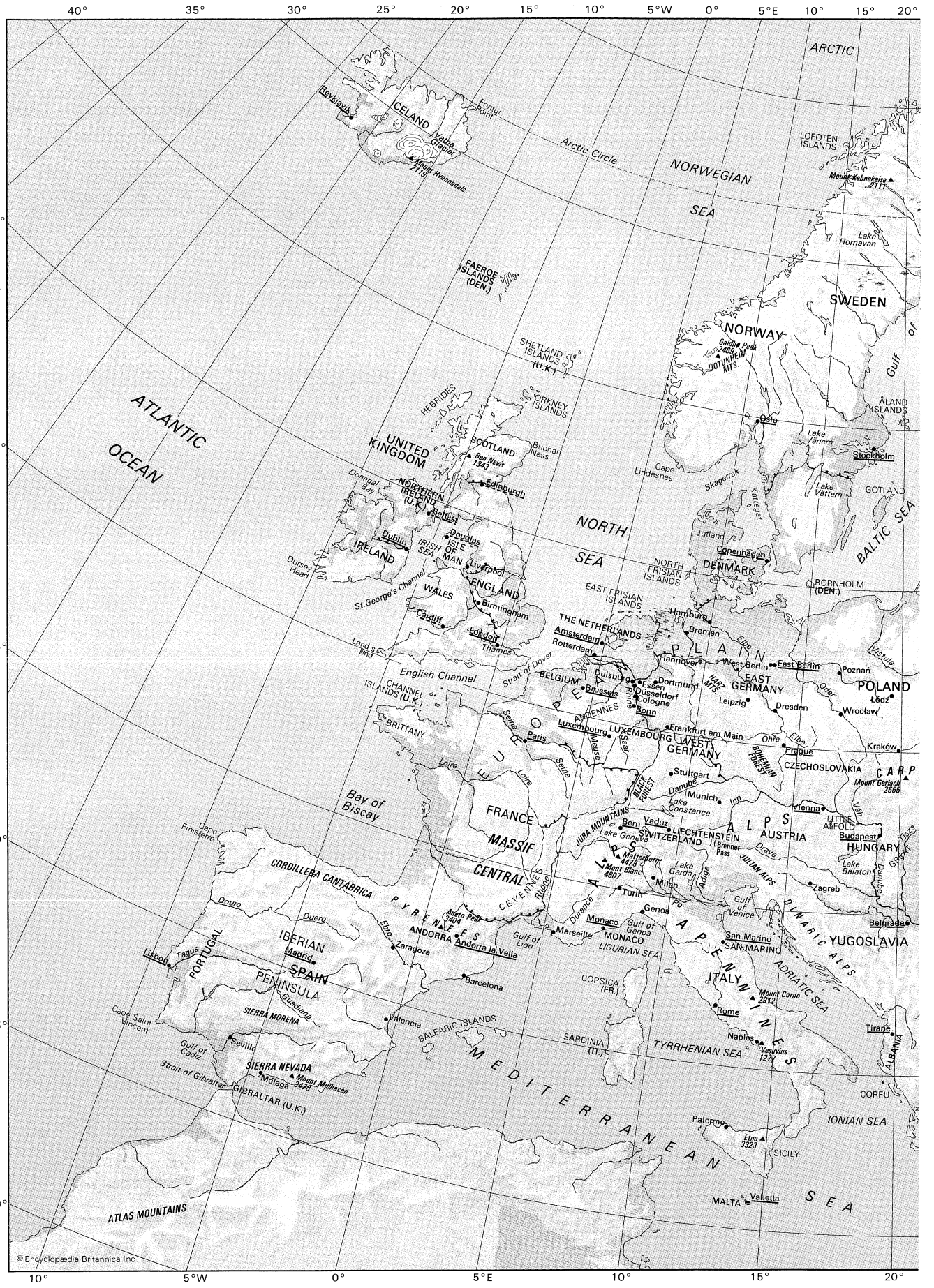
Geologic history

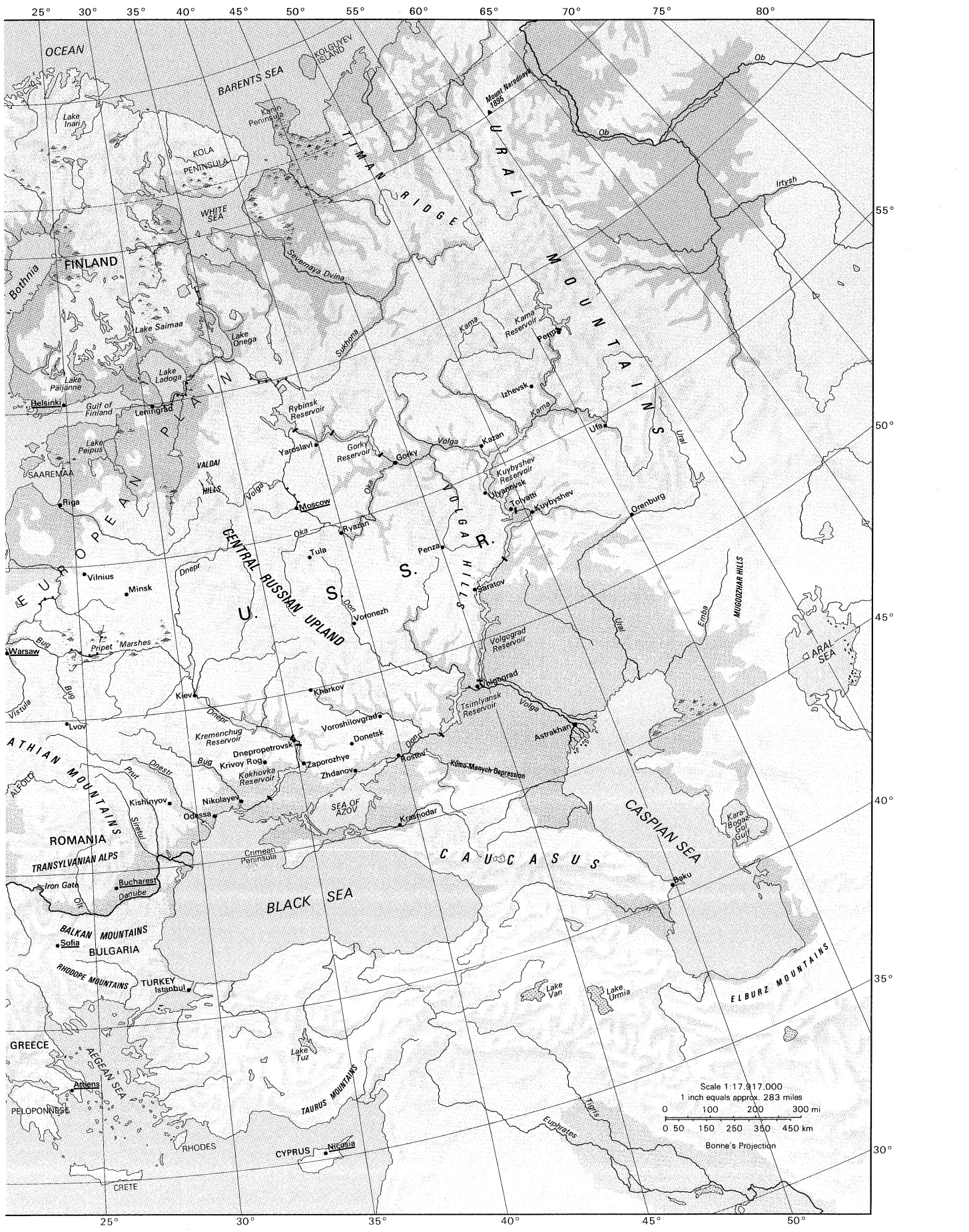
The geologic record of the continent of Europe started about three billion years ago and has continued intermittently to the present. It is a classic example of how a continent has grown through time. The Precambrian rocks in Europe range in age from about three billion to 570 million years and are succeeded by rocks of the Paleozoic era, which continued to 245 million years ago; the Mesozoic era, which lasted until 66.4 million years ago; and the Cenozoic era, which continues to today. The present shape of Europe did not finally emerge until the late Tertiary period, about five million years ago. The types of rocks, tectonic belts, and sedimentary basins that developed throughout the geologic history of Europe strongly influence human activities today.

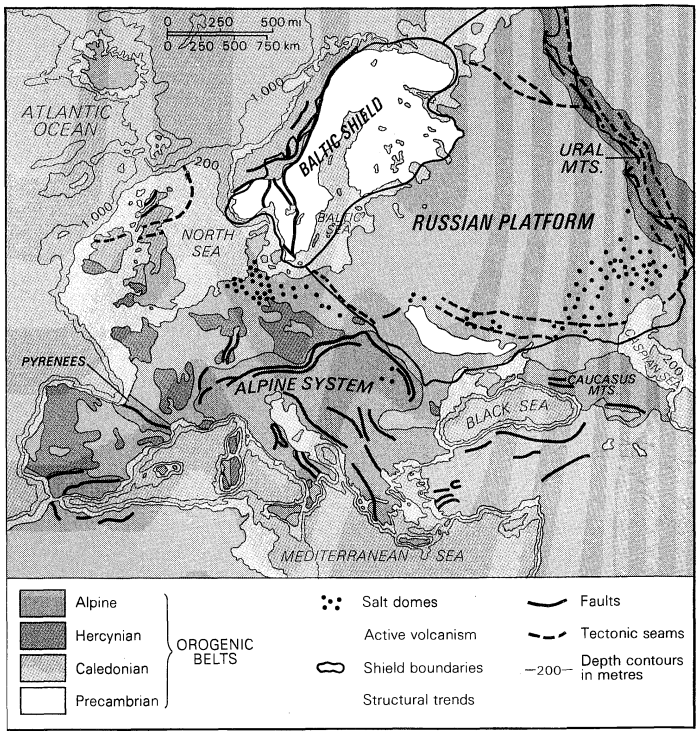
GENERAL CONSIDERATIONS

Tectonic framework. The tectonic map of Europe shows the distribution of the main tectonic units. The largest area of oldest rocks is the Baltic Shield, which has been eroded down to a low relief; the youngest rocks occur in the Alpine system, which still survives as high mountains. Between these belts are basins of sedimentary rocks that form rolling hills, as in the Paris Basin and southeastern England, or an extensive plain, as in the Russian Platform. The North Sea is a submarine sedimentary basin on the shallow-water continental margin of the Atlantic Ocean. Iceland is a unique occurrence in Europe, because it is a volcanic island situated on the Mid-Atlantic Ridge within the still-opening Atlantic Ocean.

Precambrian rocks occur in three basic tectonic environments. The first is in shields, like the Baltic Shield, which are large areas of stable Precambrian rocks usually surrounded by later orogenic belts. The second is as the basement to a younger cover of Phanerozoic sediments (*i.e.,* deposits that have been laid down since the beginning of the Paleozoic). For example, the sediments of the Russian Platform are underlain by Precambrian basement, which extends from the Baltic Shield to the Ural Mountains, and


ARCTIC
ICELAND
Reykjavik
Vatna Glacier
Mount Hvannadals 2119
Fontur Point
Arctic Circle
NORWEGIAN SEA
LOFOTEN ISLANDS
Mount Kebnekaise 2111
Lake Hornavan
SWEDEN
NORWAY
Galdhø Peak 2469
JOTUNHEIM MTS.
Gulf of
ÅLAND ISLANDS
Oslo
Lake Vänern
Stockholm
Lake Vättern
GOTLAND
BALTIC SEA
FAEROE ISLANDS (DEN.)
SHETLAND ISLANDS (U.K.)
ORKNEY ISLANDS
HEBRIDES
ATLANTIC OCEAN
UNITED KINGDOM
SCOTLAND
Ben Nevis 1343
Buchan Ness
Edinburgh
Donegal Bay
NORTHERN IRELAND (U.K.)
Belfast
Dublin
IRELAND
IRISH SEA
Douglas
ISLE OF MAN
Liverpool
Dursey Head
WALES
ENGLAND
St. George's Channel
Birmingham
Cardiff
London
Thames
Land's End
English Channel
Strait of Dover
CHANNEL ISLANDS (U.K.)
NORTH SEA
Jutland
Cape Lindesnes
Skagerrak
Kattegat
Copenhagen
DENMARK
NORTH FRISIAN ISLANDS
BORNHOLM (DEN.)
EAST FRISIAN ISLANDS
Hamburg
Bremen
Elbe
THE NETHERLANDS
Amsterdam
Rotterdam
Hannover
West Berlin
East Berlin
Poznan
Vistula
Oder
POLAND
Łódz
Wrocław
EAST GERMANY
Harz Mts.
Leipzig
Dresden
BELGIUM
Brussels
Duisburg
Essen
Dortmund
Düsseldorf
Cologne
Rhine
Bonn
ARDENNES
Luxembourg
LUXEMBOURG
Saar
Meuse
Frankfurt am Main
WEST GERMANY
Ohre
Elbe
Prague
Kraków
CZECHOSLOVAKIA
CARPATHIAN MTS.
Mount Gerlach 2655
BRITTANY
Seine
Paris
Loire
Stuttgart
Danube
BLACK FOREST
Munich
Lake Constance
Inn
Vienna
Váh
UPPER ALFOLD
Budapest
HUNGARY
Tisza
Great
Danube
Lake Balaton
Bay of Biscay
FRANCE
MASSIF CENTRAL
JURA MOUNTAINS
Bern
Vaduz
LIECHTENSTEIN
SWITZERLAND
Lake Geneva
ALPS
AUSTRIA
Brenner Pass
Drava
JULIAN ALPS
Matterhorn 4478
Mont Blanc 4807
Lake Garda
Adige
Milan
Turin
Po
Gulf of Venice
Zagreb
DINARIC ALPS
Cape Finisterre
CORDILLERA CANTÁBRICA
PYRENEES
Douro
Duero
Ebro
Aneto Peak 3404
ANDORRA
Andorra la Vella
Durance
Rhône
CÉVENNES
Monaco
Gulf of Genoa
Marseille
Gulf of Lion
LIGURIAN SEA
Genoa
APENNINES
San Marino
SAN MARINO
YUGOSLAVIA
Belgrade
ADRIATIC SEA
Lisbon
Tagus
PORTUGAL
Madrid
IBERIAN PENINSULA
SPAIN
Zaragoza
Barcelona
CORSICA (FR.)
ITALY
Mount Corno 2912
Rome
Tirane
ALBANIA
Cape Saint Vincent
Guadiana
SIERRA MORENA
Valencia
BALEARIC ISLANDS
SARDINIA (IT.)
TYRRHENIAN SEA
Naples
Vesuvius 1277
Gulf of Cadiz
Seville
SIERRA NEVADA
Mount Mulhacén 3478
Málaga
Strait of Gibraltar
GIBRALTAR (U.K.)
MEDITERRANEAN SEA
CORFU
IONIAN SEA
Palermo
Etna 3323
SICILY
MALTA
Valletta
ATLAS MOUNTAINS
© Encyclopædia Britannica Inc.
55°
50°
45°
40°
35°
30°
40°
35°
30°
25°
20°
15°
10°
5°W
0°
5°E
10°
15°
20°


25° 30° 35° 40° 45° 50° 55° 60° 65° 70° 75° 80°
55° 50° 45° 40° 35° 30°
OCEAN
BARENTS SEA
KOLGUYEV ISLAND
Kanin Peninsula
Lake Inari
KOLA PENINSULA
WHITE SEA
TIMAN RIDGE
Mount Narodnaya 1895
URAL MOUNTAINS
Ob
Irtysh
FINLAND
Bothnia
Severnaya Dvina
Lake Saimaa
Lake Onega
Sukhona
Kama
Kama Reservoir
Perm
Lake Ladoga
Lake Päijänne
Helsinki
Gulf of Finland
Leningrad
Rybinsk Reservoir
Izhevsk
Ufa
Ural
EUROPEAN PLAIN
Lake Peipus
Yaroslavl
Gorky Reservoir
Gorky
Volga
Kazan
Kuybyshev Reservoir
Ulyanovsk
Tolyatti
Kuybyshev
Orenburg
SAAREMAA
VALDAI HILLS
Riga
Volga
Moscow
Oka
Ryazan
Penza
Saratov
Vilnius
Minsk
Dnepr
CENTRAL RUSSIAN UPLAND
Tula
S. U. S. R.
VOLGA HILLS
Don
Voronezh
Volgograd Reservoir
Volgograd
Ural
Emba
MUGODZHAR HILLS
ARAL SEA
Warsaw
Bug
Pripet Marshes
Vistula
Kharkov
Kiev
Tsimlyansk Reservoir
Volga
Astrakhan
Lvov
Dnepr
Voroshilovgrad
Kremenchug Reservoir
Donetsk
Don
Dnepropetrovsk
Krivoy Rog
Zaporozhye
Rostov
Kuma-Manych Depression
CARPATHIAN MOUNTAINS
ALFOLD
Prut
Dniestr
Kakhovka Reservoir
Zhdanov
Kishinyov
Nikolayev
Odessa
SEA OF AZOV
Krasnodar
CASPIAN SEA
Kara-Bogaz-Gol Gulf
ROMANIA
Siretul
TRANSYLVANIAN ALPS
Crimean Peninsula
CAUCASUS
Iron Gate
Bucharest
Danube
Olt
Baku
BLACK SEA
BALKAN MOUNTAINS
Sofia
BULGARIA
RHODOPE MOUNTAINS
TURKEY
Istanbul
Lake Van
Lake Urmia
ELBURZ MOUNTAINS
GREECE
AEGEAN SEA
Lake Tuz
TAURUS MOUNTAINS
Tigris
Euphrates
Athens
PELOPONNESE
RHODES
CYPRUS
Nicosia
CRETE
Scale 1:17,917,000
1 inch equals approx. 283 miles
0 100 200 300 mi
0 50 150 250 350 450 km
Bonne's Projection



Structural features of Europe.

Precambrian rocks underlie the Phanerozoic sediments in southeastern England. The Ukrainian Massif is an uplifted block of Precambrian basement that rises above the surrounding plain of younger sediments. The third is as relicts in younger orogenic belts. For example, there are Precambrian rocks in the Bohemian Massif that are one billion years old and rocks in the Channel Islands in the English Channel that are 1.6 billion years old, both of which are remnants of the Middle Proterozoic era within the late Paleozoic Heryncian belt. In addition, the Precambrian Rhodope Mountains on the Balkan Peninsula are a relict block within the Alpine system.

Paleozoic sedimentary rocks either occur in sedimentary basins like the Russian Platform—which has never been affected by any periods of orogenesis and thus has sediments that are still flat-lying and fossiliferous—or occur within orogenic belts, such as the Caledonian and Hercynian, where they have commonly been deformed by folding and thrusting, partly recrystallized, and subjected to intrusion by granites.

Mesozoic–Cenozoic sediments occur either in a well-preserved state in sedimentary basins unaffected by orogenesis, as within the Russian Platform and under the North Sea, or in a highly deformed and metamorphosed state, as in the Alpine system.

Chronological summary. The geologic development of Europe may be summarized as follows. Archean rocks (those more than 2.5 billion years old) are the oldest of the Precambrian period and crop out in the northern Baltic Shield, the Ukraine, and northwestern Scotland. Two major Proterozoic orogenic belts (*i.e.,* between 2.5 billion and 570 million years old) also extend across the central and southern Baltic Shield. Thus, this shield has a composite origin, containing remnants of several Precambrian orogenic belts.

About 570 to 500 million years ago a series of new oceans opened, and their closure gave rise to the Caledonian, Hercynian, and Uralian orogenic belts. There is considerable evidence which suggests that these belts developed by plate-tectonic processes, and they each have a history that lasted several hundred million years. Formation of these belts gave rise to the supercontinent of Pangaea; its fragmentation at the beginning of the Middle Triassic epoch (about 240 million years ago) gave rise to a new ocean, the Tethys Sea. Closure of this ocean early in the Tertiary period, about 50 million years ago, by subduction and plate-tectonic processes led to formation of the Alpine orogenic system, which extends from the Atlantic to Turkey and contains many separate orogenic belts (which remain as mountain chains), including the Pyrenees, Betics, Atlas, Swiss-Austrian Alps, Apennines, Carpathians, Dinaric Alps, and Taurus and Pontic mountains. During the time that the Tethys was opening (about 180 million years ago), the Atlantic Ocean also began to open; the structure and age of the seafloor between Iceland and the continental margin of the British Isles and Norway are well known. The Atlantic is still opening along the Mid-Atlantic Ridge under the ocean, with Iceland constituting an area of the ridge that is raised above sea level. The youngest tectonic activity in Europe is represented by the present-day volcanic eruptions in Iceland; volcanoes, such as Etna and Vesuvius; and earthquakes, as in the Aegean and Turkey in the Alpine system, which result from current stresses between Europe and Africa.

Alpine orogenic system

STRATIGRAPHY AND STRUCTURE

The Precambrian. This major period of geologic time can be subdivided into the older Archean and the younger Proterozoic eons, the time boundary between them being 2.5 billion years ago. Compared with most of the other continents, Europe has few exposed Archean rocks. Some granitic gneisses, which are more than three billion years old, crop out in the northern Baltic Shield, the Ukrainian Massif, and northwestern Scotland. These rocks were recrystallized at a depth of about 12 miles (20 kilometres) in the Archean crust, but their tectonic environment is poorly understood. The Baltic Shield exhibits successively younger orogenic belts toward the south, from the Archean relicts in the north to the Late Proterozoic belt of the Sveconorwegian in southwestern Norway. A major orogenic belt, the Svecofennian, developed in the Early Proterozoic era (2.5 to 1.6 billion years ago); it now occupies the bulk of the Baltic Shield, especially in Finland and Sweden, where it extends from the Kola Peninsula to the Gulf of Finland near Helsinki. The Sveconorwegian is a north–south-trending orogenic belt that developed in the period between 1.2 billion and 850 million years ago. It occupies southern Norway and the adjacent area of southwestern Sweden between Oslo and Göteborg. On its northern side it has been reactivated almost beyond recognition within the Caledonian orogenic belt. The Ukrainian Massif and the small Laxfordian belt in northwestern Scotland consist mainly of granitic rocks and highly deformed and metamorphosed schists and gneisses that originally were sediments and volcanics, and their age is similar to that of the Svecofennian belt. In northwestern Scotland there is a north–south-trending belt of red sandstones and conglomerates belonging to the Torridonian group that is about one billion years old; these sediments may be the erosional products or molasse of a 1.2-billion-year-old orogenic belt, of which there are a few relicts within the Paleozoic Caledonian belt of Scotland. The Bohemian Massif is a diamond-shaped block in the heart of Europe, which has been heavily affected by the late Paleozoic Hercynian orogeny. Many of the rocks originally formed in the Archean (about 2.7 billion years ago), the Early Proterozoic (Svecofennian times), or later in the Proterozoic (about one billion years ago), however, were strongly deformed in several Precambrian orogenies and thus are now schists, gneisses, and amphibolites, accompanied by a variety of granites. Near the end of the Precambrian—about 800 to 570 million years ago—there was widespread deposition of conglomerates, sandstones, clays, and some volcanic sediments, which make up the Eocambrian (or Vendian) group; these were derived from the erosion of uplifted Precambrian mountains. They are well known for two features. First is their glacial sediments, which were deposited at a time of worldwide glaciation; they occur in northwestern Scotland (Islay Island), western Ireland, Norway (Finnmark and West Spitzbergen), Sweden, France (Normandy), and Czechoslovakia (Bohemian Massif). Second is the occurrence of impressions of soft-bodied organisms, such as seaweed, jellyfish, and worms, which represent the beginnings of Metazoan life before the explosion of life forms with hard parts for skeletons that became abundant in the Early

Cambrian. These impressions occur in Charnwood Forest in central England, southern Wales, northern Sweden, the Ukraine, and several localities in the Russian Platform.

Mineral deposits

The Precambrian rocks of Europe provide a rich source of economic minerals to sustain human activities, such as major iron ore deposits at Kiruna in northern Sweden and Krivoy Rog in the Ukraine; tin deposits associated with granites in Finland; extensive copper–nickel sulphide ores across Finland, especially at Outokumpu, and in Sweden; and magnetite ores containing vanadium and titanium in northern Finland.

The Paleozoic era. The Paleozoic (570 to 245 million years ago) tectonic geology of Europe can be divided into two parts: the major orogenic belts of the Caledonian (or Caledonides), the Hercynian (or Hercynides), and the Uralian (or Uralides); and the undisturbed, mostly subsurface (and thus poorly known) Paleozoic sediments in the triangular area between these belts in the Russian Platform.

Caledonian orogenic belt. The major factor that controlled the early mid-Paleozoic development of Europe was the opening and closing of the Iapetus Sea, which gave rise to the Caledonian orogenic belt that extends from Ireland and Wales through northern England and Scotland to western Norway and northward to Finnmark in northern Norway. The belt is confined between the stable blocks of the Baltic Shield and the Precambrian belt of northwestern Scotland. Remnants of the Iapetus seafloor are seen in ophiolites at Ballantrae in Strathclyde region, Scot., and near Bergen, Nor. During the Cambrian period (570 to 505 million years ago) widening of the Iapetus gave rise to extensive shelf seas on the bordering continents, which deposited a thin cover of limestone and shale with a remarkable diversity of fossils of numerous marine invertebrates. The presence of this sea can be demonstrated by the fact that trilobites and graptolites in northern Scotland, which was on one side, are significantly different from those in central England and southern Norway, which were on the other. In the Ordovician period (505 to 438 million years ago) the sea began to close by subduction, giving rise to major magmatic belts with lavas and tuffs in the Lake District of northern England and Snowdonia National Park in Wales—where there is associated gold and copper mineralization—and to many granites in the highlands of Scotland. In the Silurian period (438 to 408 million years ago) the Iapetus Sea closed, with the result that the bordering continental blocks collided, giving rise to deformation, metamorphism, and the orogeny of the Caledonian belt. In the Late Silurian, early land plants and the first freshwater fish appeared in lakes on the belt. The rifts of the Orkney Basin developed in the Devonian period (408 to 360 million years ago) on top of the thickened and unstable crust of the Caledonian orogenic belt in a manner comparable to the Quaternary rifts of Tibet (*i.e.,* those that have appeared in the last 1.6 million years) that has a crust thickened by the Himalayan orogeny of the Tertiary period (66.4 to 1.6 million years ago). Erosion of the uplifted mountain belt in the Devonian led to deposition of sandstones and conglomerates in basins over a wide region from the British Isles to the western Russian Platform, often called the Old Red Sandstone continent.

Hercynian orogenic belt. The Hercynian, or Variscan, orogenic belt evolved during Devonian and Carboniferous times, from about 408 to 286 million years ago. The belt extends form Portugal and western Spain, southwestern Ireland, and southwestern England in the west; through the Ardennes, France (Brittany, Massif Central, Vosges, and Corsica), Sardinia, West and East Germany (Odenwald, Black Forest, and Harz mountains); to Czechoslovakia (Bohemian Massif). The orogeny was formed by plate-tectonic processes that included sea-floor spreading, continental drift, and the collision of plates. Remnants of the original ocean floor are preserved as ophiolites in the Harz Mountains and in the Lizard Peninsula of southwestern England. In the Devonian a continental margin ran along the north side of the belt in Devon and Cornwall (England) on which extensive sandstones derived from the continent were deposited. In the Carboniferous period, shallow-water limestones were laid down in the area of the Pennines of England on a shelf or carbonate bank; this formation passes southward into deeper-water shales of the Culm Trench of southwestern England, within which are found the pillow lavas, gabbros, and serpentinites of the Lizard ophiolite. In Brittany there is an island arc with lavas and granites that resulted from subduction of the ocean floor. The main Hercynian suture zone of the collided plates extends from the south side of Brittany to the Massif Central. Throughout much of Europe there is evidence of extensive thrusting, implying that there was appreciable thickening of the continental crust and the formation of a Tibetan-style plateau across the Hercynian orogeny. The thickening led to melting of the lower crust and the formation of large numbers of Late Carboniferous granites, especially in the Massif Central. The plateau became overly thick and unstable, and this caused the formation of rifts that developed into coal-bearing basins—as in Silesia (Poland) and the Massif Central—in the Late Carboniferous and Permian periods (*i.e.,* between about 320 and 245 million years ago). The varied tectonic development of the Hercynian orogeny gave rise to widespread mineral deposits in many environments, which have been exploited in the economic development of many countries. Lead and zinc deposits occur in shelf carbonate sediments in Ireland and the Pennines of England; there are deposits of copper, lead, and zinc sulfides that formed in rifts in Silesia (Poland and East Germany) and at the Riotinto Mines in southwestern Spain; and important mineral deposits of tin, tungsten, and uranium are associated with crustal melt granites in Cornwall, the Massif Central, and Spain and Portugal.

Uralian orogenic belt. The Uralian orogenic belt extends for about 1,550 miles (2,500 kilometres) from the Arctic Ocean to the Aral Sea and forms the traditional eastern boundary of Europe. This is a late Paleozoic belt that developed as a result of collision between Asia and Europe. The earliest rifts in old Precambrian basement rocks began in the Late Cambrian–Early Ordovician, about 500 million years ago, and these developed into the floor of a new ocean. Island arcs formed in the Silurian period and countless ophiolitic slabs of ocean floor were thrust onto the continental margins. In Devonian times a considerable amount of thrusting and metamorphism occurred, and the final parts of the ocean floor were subducted (*i.e.,* thrust under continental masses); the result of this activity was that in the Carboniferous there was a final collision between the continents of Europe and Asia that gave rise to the Uralian orogenic belt. In the Permian there was widespread deposition of limestones followed by red sandstones, which were derived by erosion of the mountains. In the 1840s the British geologist Sir Roderick Murchison coined the term Permian System for the city of Perm. The Ural Mountains are very rich in mineral deposits—especially chromite, platinum, nickel, copper, and gold—which are associated with the major ophiolitic slabs of ocean floor distributed along the mountain chain.

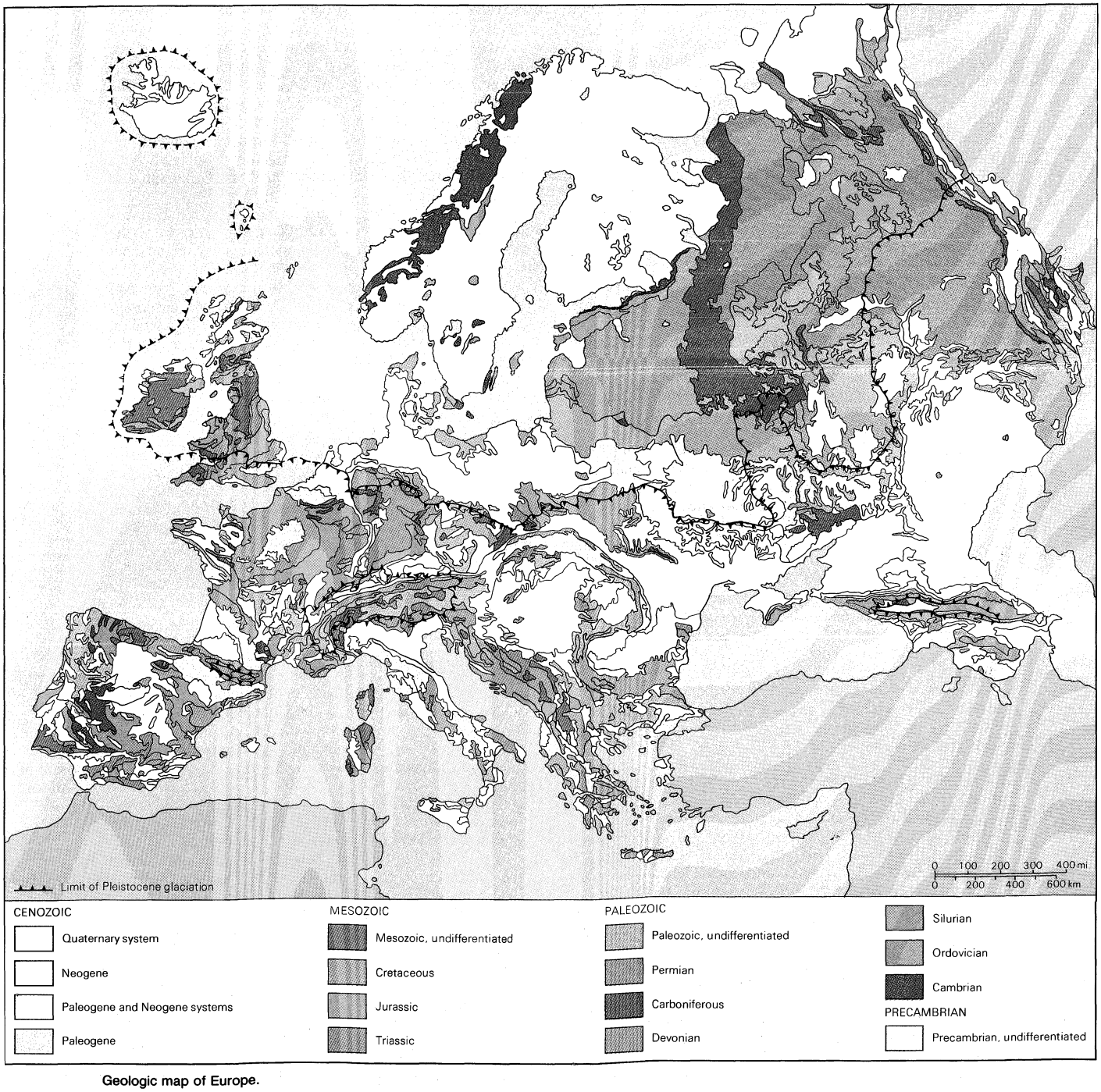
The Mesozoic and Cenozoic eras. During the Mesozoic era a new ocean, the Tethys, evolved in what is now southern Europe, and during the Cenozoic era this ocean was destroyed by subduction with the result that many small plates collided. These events gave rise to the present-day tectonic mosaic that extends eastward from the Atlas Mountains of North Africa, the Betic Cordillera of southern Spain, and the Pyrenees; via the Alps of maritime France, Switzerland, and Austria; to the Carpathians, the Apennines, the Dinaric Alps, the Alpine belt of Bulgaria, the Taurus and Pontic mountains of Turkey, and finally to the Caucasus. Within these belts must also be included the Pannonian Basin of Romania and the Balearic, Alboran, Tyrrhenian, and Adriatic basins of the Mediterranean Sea. The main cause of this Alpine orogeny during the Cenozoic was the northward compression of Africa into Europe.

Tectonic mosaic of southern Europe

The first rifting of the older continent began with salt and evaporite deposition in lakes in rift valleys in the Early Triassic (245 to 240 million years ago). By 220 million years ago, in the Late Triassic, the continental margins of the new, narrow Tethys were commonly covered by shallow water over fossiliferous, carbonate shelf sediments.

During the Middle Jurassic, about 180 million years ago, these carbonate shelves began to fragment, and in the Cretaceous (144 to 66.4 million years ago) the ocean floor was subducted in many places. This gave rise to volcanic island arcs, such as those of present-day Indonesia, and slabs of the Tethys ocean floor were thrusted as ophiolites onto the continental margins. Extensive remnants of these ophiolites can be seen today, especially in the northern Apennines and in Yugoslavia, Greece, Turkey, and Cyprus. Collisions between many of the continental microplates took place in the Eocene–Oligocene (about 58 to 24 million years ago) epochs. For example, the Iberian Peninsula rotated to give rise to the Pyrenees, the Italian Peninsula drove northward and compressed into Europe, causing growth of the Swiss-Austrian Alps, and Anatolia moved westward and gave rise to the Aegean arc and the mountains of Greece. It is interesting to consider that it was the opening of the Red Sea that caused the Arabian Peninsula to slide northward along the fault defined by the Dead Sea and the Jordan Valley and in so doing to form at its front the Zagros Mountains of Iran, which, in turn, pushed Anatolia westward and caused the deformation in Greece. This scenario illustrates the interlinking and interdependence of all these movements and structures in Europe with those outside the continent. In the Late Miocene (11.2 to 5.3 million years ago) many of the early Mediterranean basins (*e.g.*, Balearic, Tyrrhenian, Ionian, and Levantine) became isolated from the main Atlantic and Indo-Pacific oceans, and in these basins were laid down huge deposits of salt and gypsum in evaporites up to more than a mile thick. There are several important economic mineral deposits in the European Alpine system that can be related to the several stages of geologic evolution described above. Lead and zinc deposits occur in Triassic shelf limestones at Blei-berg, W.Ger. There are many chromite ores in the ophiolites of Yugoslavia, Greece, and Turkey. Copper ores formed in pillow-bearing basaltic lavas of the Tethyan ocean floor; copper mines

Mineral deposits in the Alps



Geologic map of Europe.

have been worked since antiquity in Cyprus, which lent its name to this element. The Tethys, however, was a relatively narrow ocean, and thus its limited subduction was not able to give rise, for example, to many granites and volcanic rocks, which might have contained useful mineral deposits. Active seismic disturbances expressed as earthquakes are a reflection of the continuing compression between several of the European microplates; they are common in the Atlas Mountains, the island arc of the South Aegean, Greece, the island arc of the Tyrrhenian Sea in southern Italy, Turkey, and the Caucasus Mountains.

The North European and Russian platforms. An approximately triangular area is described between the Caledonian orogeny in the west, the Hercynian orogeny and the Alps in the south, and the Urals in the east. This area includes the Russian and North European platforms and the North Sea. Within this area the Phanerozoic sedimentary rocks are either undeformed or only weakly deformed, and thus this area contrasts with the surrounding orogenic belts described above where such sediments are strongly deformed. Thus, throughout much of the extensive Russian Platform the Paleozoic, Mesozoic, and Cenozoic sediments have escaped the effects of the surrounding orogenies, and they are almost as horizontal as when they were laid down. Farther west in the portion of the North European Platform that includes southeastern England and northern France, Mesozoic and early Cenozoic sediments have been weakly deformed into anticlines and synclines by the Tertiary deformation of the Alpine orogenic belt to the south. This took place at a shallow level of the crust, and the sediments are still unmetamorphosed. Thus, the best place to find beautifully preserved Phanerozoic fossils is in this central triangular area of Europe. Under the North Sea there are gas reserves in Permian and Triassic sediments, and there are major oil reservoirs in Jurassic sediments. This is a subsided fragment of the continental margin of Europe flooded with water from the melted glaciers of the last Ice Age.

The Tertiary igneous province of northwestern Britain. From about 61 to 52 million years ago (early in the Tertiary) there were important igneous extrusions and intrusions in northwestern Britain. In Northern Ireland and northwestern Scotland, basaltic lava flows (*e.g.,* the Giant's Causeway and the northern part of the isle of Skye) are associated with northwest–southeast-trending basaltic dikes and many plutonic complexes, which are probably the roots of volcanoes. The dikes extend southeastward across northern England and continue under the North Sea. Related lavas occur in the Faeroe Islands, belonging to Denmark. These igneous rocks formed in the faulted and thinned continental margin of northwestern Europe contemporaneously with the rifting and seafloor spreading that gave rise to the Atlantic Ocean.

Basaltic lava flows and dikes

Iceland. The Mid-Atlantic Ridge is a major plate boundary separating the North American and the Eurasian plates, and it extends through the centre of Iceland. Along this ridge the Atlantic Ocean is still growing, and on Iceland this activity is expressed as major rifts, volcanoes, and steam geysers. The entire island is made of lavas, the oldest of which on the northwestern coast came from eruptions about 16 million years ago. Iceland thus preserves a unique record of the last stages of development of one of the world's major accreting plate boundaries, most of which is elsewhere submarine. (B.F.W.)

The Quaternary period. The Pleistocene epoch occupies the Quaternary period (the last 1.6 million years), with the exception of the last 10,000 years, which are called the Holocene epoch. Although the precise causes of the Ice Ages that mark the Pleistocene are controversial, it is known that prior to this glaciation northern Europe had risen to a much higher elevation than now and that ice formed to great depths there, as in the rest of the Atlantic landmass and the Alpine areas. The Pleistocene was punctuated by warm interglacial periods separating glacial advances; during its latter part, humans occupied niches in the more southerly parts of the continent.

Glaciers are the most powerful engines provided by nature for the transport—by plucking or quarrying—of large masses of rock, and certainly the European glaciers transformed the physique both of their source areas and of the lands to which they moved. Many physical forms of northern and Alpine Europe resulted from glacial erosion, supplemented by weathering, and the surfaces of areas where the glaciers eventually withered away consisted of masses of transported material. Southern Scandinavia, southern Finland, the Swiss Plateau, and the North European Plain were thickly plastered with a variety of forms, including boulder-studded clay, gravels, sands, and the windblown deposits known as loess. New drainage patterns were formed. The melting of so much ice raised the level of the oceans by an estimated 320 or more feet, while former ice-clad lands, including the North Sea area, began to rise isostatically. It was not until quite late in the Holocene that the northern seas of Europe—the Irish, North, and Baltic—took, by stages, their present shape.

The modern geologic framework. Although the exposed rocks of Europe are obscured increasingly by the works of humans, and while detailed understanding of rock patterns present challenges even to the expert, the major formations of the continent are clear. In the north lie wide areas of ancient worn-down rocks, stripped of soil by the glaciers but compensated in some measure by the coastal plains created by uplift. In contrast, southern Europe, although incorporating such relicts as massifs of Paleozoic rocks, is essentially a youthful world, not yet fully fashioned, as evidenced by continuing seismic disturbances. Eastern Europe, based on the vast Russian Platform, is a stable world still young in surface, since the floor of its shield rocks is deeply concealed beneath Mesozoic and Tertiary deposits, above which glacial material covers the northern half and loess deposits enrich the south. Although in scale this platform is a continental area, river development facilitates access to inland seas in both the north and the south. Ancient rocks, lying near the surface, offer mineral wealth, and the former Volga–Ural seas have left a residue of petroleum and mineral salts. For the rest, western and central Europe show great diversity of landforms and landscape as well as varied soil and mineral resources. Alpine ranges in the south and southeast combine high altitude and relief with scenic attractions and—more importantly—with high precipitation and water dispersion. Highland areas, remnants of faulted Hercynian belts surrounded by younger strata, provide another type of wooded landscape, with their contained coalfields. Iceland has the youngest landscape of Europe, with its spectacular semiactive volcanoes, high waterfalls, extensive glaciers, and steam geysers. Lastly, lowlands, of great human value, recall their varied origins—former sea and lake basins; lowlands of glacial deposition; parts of eroded synclinal structures; and alluvial and marine plains won from the sea by isostasy or, as exemplified by the Dutch polders, by the work of humans. (W.G.E./B.F.W.)

Differences between northern and southern Europe

The land

A contrast exists between the configuration of peninsular (or western) Europe, and eastern Europe, which is a much larger and more continental area. A convenient division is made by a line linking the base of the Jutland Peninsula with the head of the Adriatic Sea. The western part of the continent clearly has a high proportion of coastline with good maritime access and often with inland penetration by means of navigable rivers. Continental shelves—former land surfaces that have been covered by shallow seas—are a feature of peninsular Europe, while the coasts themselves are both submerged or drowned, as in southwestern Ireland and northwestern Spain, and emergent, as in western Scotland and southern Wales where raised former beaches are in evidence. East of the Vistula River, Europe's expansive lowlands have something of the scale and character of those of northern Asia, but the continent also comprises numerous islands, some—notably the Faeroes and Iceland—located at a distance from the mainland. Fortuitously, Europe has no continuous mountain obstacle aligned north–south, corresponding, for example, to the Western Cordillera of North America and the Andes of South America, that would limit access into western Europe from the ocean.

Peninsular and eastern Europe

RELIEF

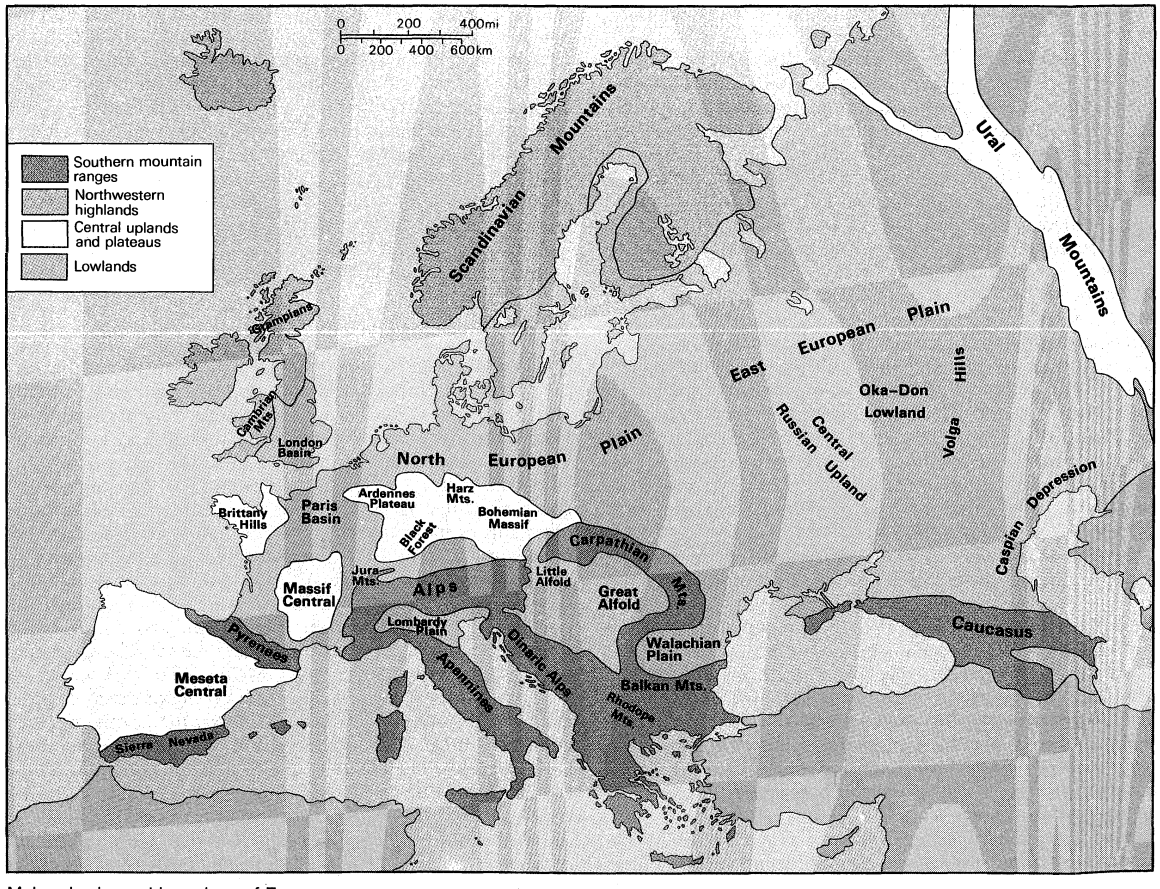
Elevations. Lands lying at high altitude can, of course, be lands of low relief, but on the European continent relief tends to become more rugged as altitude increases. The greater part of Europe, however, combines low altitude with low relief. Only hill masses less than 800 feet (240 metres) in height rise gently within the Russian (or East European) Plain, which continues northward into Finland, westward into the North European Plain, and southward in the Romanian, Bulgarian, and Hungarian plains. The North European Plain, common to much of Poland, northern Germany, and Denmark, broadens in western France and continues, across the narrow seas, in southeastern Great Britain and Ireland. The major peninsula of Scandinavia is mostly upland and highland, with its relief greatest at the descent to the Norwegian fjords and the sea; eastward and southward the seas are approached more gently. The highest points reached in Norway and Sweden are, respectively, Galdhø Peak (8,100 feet) and Mount Kebne (6,926 feet). Iceland's highest peak is Mount Hvannadals, at 6,952 feet, while Ben Nevis, the highest summit in Great Britain, stands at a height of only 4,406 feet. Greater relief is found in those areas in the heart of western and central Europe where uplifted and faulted massifs survive from the Hercynian orogeny. The worn-down Ural Mountains also belong in this category, and their highest point, Mount Narodnaya (6,217 feet), corresponds approximately to that of the Massif Central in south central France. Altitudes in these areas are mainly between about 500 and 2,000 feet, and many steep slopes are to be seen.

The highest altitudes and the most rugged relief of the European continent are found farther south, where the structures of the Cenozoic orogeny provide mountain scenery. In the Alps, Mont Blanc rises to a height of 15,771 feet (4,807 metres), which is the highest point on the continent. In the Pyrenees and the Sierra Nevada of Spain, the highest of the peaks exceed 11,000 feet. The Apennines, Dinaric Alps, and Balkan Mountains, as well as the arc-shaped Carpathians and their southern portion, the Transylvanian Alps, also exhibit high altitudes. The highest peaks in these ranges are Mount Corno (9,554 feet) in the Abruzzi Apennines, Bobotov Kuk (8,274 feet) in the Dinaric Alps, Mount Botev (7,795 feet) in the Balkan Mountains, Gerlachovský Štít (Gerlach; 8,711 feet) in the Western Carpathians, and Mount Moldoveanu (8,347 feet) in the Transylvanian Alps. Above all, in southern Europe—Austria and Switzerland included—level, low-lying land is scarce, and mountain, plateau, and hill landforms dominate. The lowest terrain in Europe, virtually lacking relief, stands at the head of the Caspian Sea; there the Caspian Depression reaches some 95 feet (29 metres) below sea level.

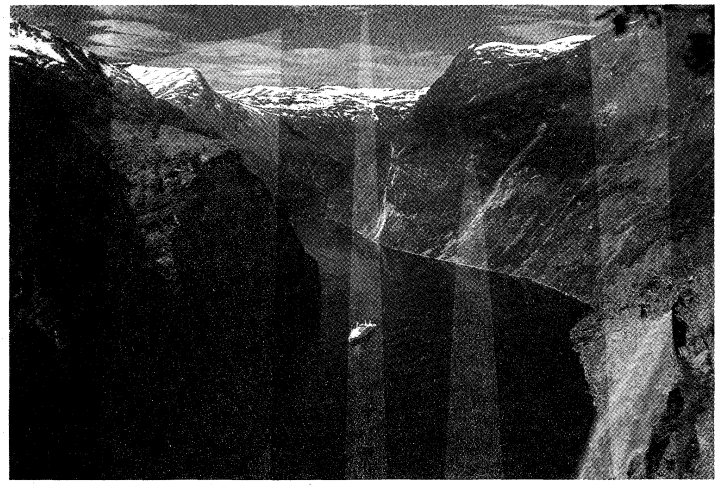
Physiographic units. Four broad topographic units can be simply, yet usefully, distinguished in the continent of Europe: coastal and interior lowlands, central uplands and plateaus, the northwestern highlands, and southern Europe.

Lowlands. More than half of Europe consists of lowlands, standing mostly below 600 feet but, infrequently, rising to 1,000 feet. Most extensive between the Baltic and White seas in the north and the Black, Azov, and Caspian seas in the south, the lowland narrows westward, lying to the south of the northwestern highlands; it is divided also by the English Channel and the mountains and plateaus of central Europe. The Danubian and northern Italian lowlands are thus mountain-ringed islands. The northern lowlands are areas of glacial deposition and, accordingly, their surface is diversified by such hills as the Valdai and the North German Heights; by deposits of boulder clay, sands, and gravels; by glacial lakes; and by the Pripet Marshes, a large ill-drained area of the western Soviet Union. Another important physical feature is the southeast–northwest zone of windblown loess deposits that have accumulated from eastern Britain to the Ukraine. This *Börde* (German: "edge") belt lies at the northern foot of the Central European Uplands and the Carpathians. Southward of the limits of the northern ice sheets are vales and hills, with the Paris and London basins typical examples. Superficial rock cover, altitude, drainage, and

The northern lowlands



Major physiographic regions of Europe.



Scenic fjord, or sea inlet, winding deep into the mountainous coast of western Norway.
Bob and Ira Spring

soil have sharply differentiated these lowlands—which are of prime importance to human settlement—into areas of marsh or fen, clay vales, sand and gravel heaths, or river terraces and fertile plains.

Central uplands and plateaus. The central uplands and plateaus present distinctive landscapes of rounded summits, steep slopes, valleys, and depressions. Examples of such physiographic features can be found in the Southern Uplands of Scotland, the Massif Central of France, the Meseta Central of Spain, and the Bohemian Massif. Routes detour around, or seek gaps through, these uplands—whose German appellation, *Horst* ("thicket"), recalls their still wooded character, while their coal basins give them great economic importance. The well-watered plateaus give rise to many rivers and are well adapted to pastoral farming. Volcanic rocks add to the diversity of these regions.

Northwestern highlands. The ancient, often mineral-laden rocks of the northwestern highlands, their contours softened by prolonged erosion and glaciation, are found throughout much of Iceland, Ireland, and in northern and western Britain and Scandinavia. These highland areas include lands of abundant rainfall—which supplies hydroelectricity and water to industrial cities—and provide summer pastures for cattle. The land in these areas, however, is of little use for crops. The coasts of the northwestern highlands—and in particular the fjords of Norway—invite maritime enterprise.

Southern Europe. A world of peninsulas and islands, southern Europe is subject to its own climatic regime, with fragmented but predominantly mountain and plateau landscapes. Iberia and Anatolia (Turkey) are extensive peninsulas with interior tablelands of Paleozoic rocks that are flanked by mountain ranges of Alpine type. The restricted lowlands lie within interior basins or fringe the coasts; those of Portugal, Macedonia, Thrace, and northern Italy are relatively large. Runoff from the Alps furnishes much water for electricity-generating stations, as well as for the flow regimes of major rivers.

Peninsulas and islands of the south

Detailed discussion of the Alps, Apennines, Carpathian Mountains, European Plain, Pyrenees, and Ural Mountains can be found in *European geographic features of special interest* at the end of this article.

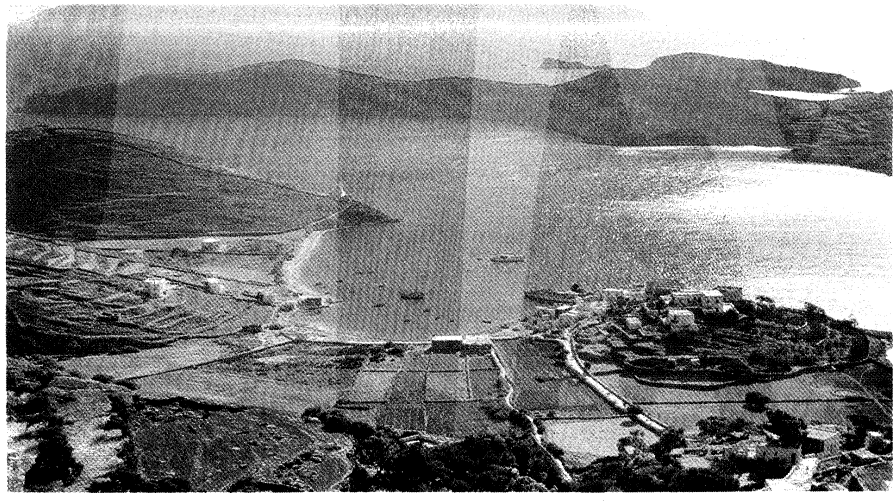
DRAINAGE

Topographic influences. The drainage basins of most European rivers lie in areas originally uplifted by the Caledonian, Hercynian, and Alpine mountain-building periods that receive heavy precipitation, including snow. Some streams, notably in the European Soviet Union, Finland, and Poland, have their sources in hills of Tertiary rocks, while others, including the Thames and Seine rivers, derive from hill country of Mesozoic rocks. Drainage is directly, or via the Baltic and the Mediterranean seas, to the Atlantic and the Arctic oceans and to the enclosed Caspian Sea.

Drainage basins

The present courses and valley forms of the major rivers result from an intricate history involving such processes as erosion by the headstream, downcutting, capture of other rivers, faulting, and isostatic changes of land and sea levels. The Rhine, for example, once drained to the Mediterranean before being diverted to its present

Josef Muench



Coastal landscape of submerged mountains and resulting islands and bays characteristic of Greece along the Aegean Sea.

northerly course. The courses of many rivers—notably those of Scandinavia and the North European Plain—have been shaped since the Pleistocene epoch. While the Alps, Apennines, and Carpathians provide watersheds, other mountain ranges have been cut through by rivers, as by the Danube at Vienna, Budapest, and the Iron Gate and by the Olt (in Romania). In the Russian Plain the rivers are long and flow sluggishly to five seas. In western, central, and eastern Europe, rivers are largely "mature"; *i.e.,* their valleys are graded, and their streams are navigable. Northern and southern Europe, in contrast, present still "youthful" rivers, as yet ill-graded and thus more useful for hydroelectricity than for waterways. The Atlantic rivers have scoured estuaries widening seaward, while, in the Baltic, Mediterranean, and Black seas, with minimum tidal influences, deltas and spits have been created. The upper Dnepr, since post-Pleistocene times, has failed to drain effectively the low area of minimal relief known as the Pripet Marshes.

Hydrography. The water volume of, and discharge from, the rivers of Europe are governed by factors that include local conditions of rainfall, snowmelt, and rock porosity. In consequence, the rivers in the western area have more volume and higher discharges in the winter season and are at their lowest in summer. In areas of mountainous and continental climate, thanks to the runoff of snowmelt, the rivers are highest in spring and early summer. The longer rivers of the continent, notably the Rhine and the Danube, have complex regimes, since their basins extend into areas of contrasting climate. Although embanking measures have reduced the problem, flooding is a continued threat. Thus, rivers in the Soviet Union are liable to flood with the spring thaw; oceanic rivers, after heavy or prolonged rain over the whole basin; and Alpine rivers, when the warm foehn wind rapidly melts the snow. In the Mediterranean region some rivers—as in peninsular Greece—tend to dry up in summer through a combination of scant rainfall, evaporation, and porous limestone beds. In the Abruzzi region of central Italy, however, heavy rainfall, mainly in winter, permeable and porous rocks within the basin, and abundant snow combine to regulate the river regimes.

Flood danger

The Rhône achieves a steady flow throughout the year, deriving a high input from the Cévennes Mountains—which experience heavy winter rain—plus abundant spring and summer snowmelt from the Alps via Lake Geneva. The Rhine and Danube tap supplies from the Alps in spring and summer, and the Rhine, especially, taps areas of winter rainfall maximum. The Volga has its highest water in spring and early summer, thanks to snowmelt, and falls to a summer low. The Saône, lying within the oceanic climatic area, tends to have a good flow year-round. The winter freeze of the east only rarely seriously affects the Danube and western European rivers.

Lake systems and marshes. Lakes cover less than 2 percent of Europe's surface and occur mostly in areas subjected to Pleistocene glaciation. The Scandinavian Peninsula and the North European Plain account for four-fifths of the area of lakes; and in Finland lakes cover one-fifth of the surface. The other major zones of lakes lie marginal to the Alpine system, while Scotland, too, has its many "lochs" and Ireland its "loughs." Lakes survive where the inflow of water exceeds loss from evaporation and outflow and should eventually disappear through alluvial accumulation. Their origins lie in the glacial excavation of softer rocks, in the building of dams by morainic material, and in tectonic, or deforming, forces, which may create depressions. This second explanation clearly applies to Alpine lakes; to many of those in the British Isles, including the small but scenic ones of the Lake District of England; and also to those of central Sweden. Volcanic crater lakes are found in central Italy, and small lakes of the lagoon type are found along the Baltic and Mediterranean, where spits have lengthened parallel to the coast and hence cut off sea access.

A well-developed zone (the *Marschen*) has formed along the low-lying and reclaimed marshes along the North Sea in West Germany and The Netherlands, and characteristically the estuaries of Europe's tidal rivers are edged by flat alluvial marshes. Fens, as exemplified by the polders in The Netherlands and the lowlands in eastern England, are made up of either alluvium or peat and stand too low to be drained effectively, except by continuous pumping. The continent's largest marshland is the Pripet Marshes of the Belorussian and Ukrainian Soviet Socialist republics.

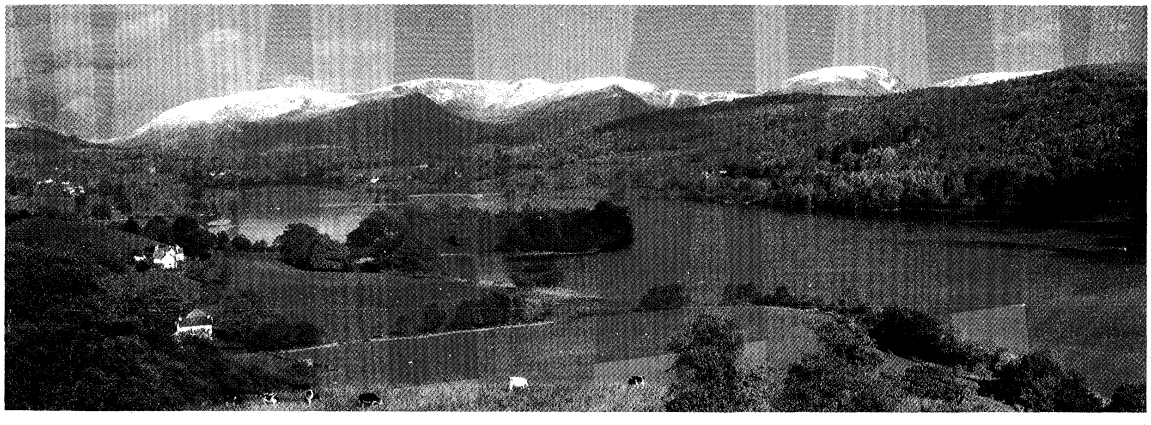
The *Marschen* zone

Detailed discussion of Europe's drainage systems can be found in *European geographic features of special interest* at the end of this article. The discussion for western Europe includes the Rhine, Rhône, and Seine rivers. The discussion for central Europe includes the Danube, Elbe, Oder, and Vistula rivers. The discussion for eastern Europe includes the Dnepr, Don, and Volga rivers.

SOILS

Regional divisions. The soil patterns of Europe are clearly and zonally arranged in the Russian Plain but are much more complicated in the rest of the continent, which exhibits a more varied geology and relief. Tundra soils occur only in Iceland, the most northerly parts of the Soviet Union and Finland, and in high areas of Sweden and Norway; they tend to be acidic, waterlogged, and poor in plant nutrients. South of this zone, and extending around the Gulf of Bothnia and across Finland and the Soviet Union north of the upper Volga, cool-climate podzols are characteristic. These soils, formed in a coniferous woodland setting, suffer from acidity, the leaching of minerals, hardpan formation and permafrost beneath the topsoil, and excess of moisture; given the climate, they are virtually useless for crops. The larger zone to the south stretches from the central Soviet Union westward to Great Britain and Ireland and southward from central Sweden, southern Norway, and Finland to the Pyrenees, Alps, and Balkan Mountains. In this region temperate-climate podzols and brown forest soils have developed in a mixed-forest environment, and these soils, which are highly varied, usually have a good humus content. Locally, the farmer recognizes soils of heavy to light texture, their different water-holding capacities, depth, alkalinity

FPG, Tom Wright



Mountain-encircled Esthwaite Water in the Lake District of northwestern England.

or acidity, and their suitability for specific crops. The soils, rich in humus, within this zone that cover loess are excellent loams; lowland clays, when broken down, also exhibit high quality, as do alluvial soils; in contrast, areas covered with dry, sandy, or gravelly soils are more useful for residential and amenity purposes than for farming. In the southwestern European Soviet Union, notably in the Ukraine, some soils that have been formed in areas of grass steppe are chernozems (black earths)—deep, friable, humus-rich, and renowned for their fertility. In the formerly wooded steppe lying to the north of the grass steppe in both the south central Soviet Union and the lower Danubian lowlands, soils of somewhat less value are known as degraded chernozems and gray forest soils. At best, chestnut soils—some needing only water to be productive—and, at worst, solonetzic (highly saline) soils cover areas of increasing aridity eastward of the Ukraine to the Ural River. Lastly, in southern Europe, where the countryside is fragmented by mountains, plateaus, and hills, much soil has been lost from sloping ground through forest destruction and erosion, and a bright red soil (terra rossa), heavy and clay-rich, is found in many valleys and depressions.

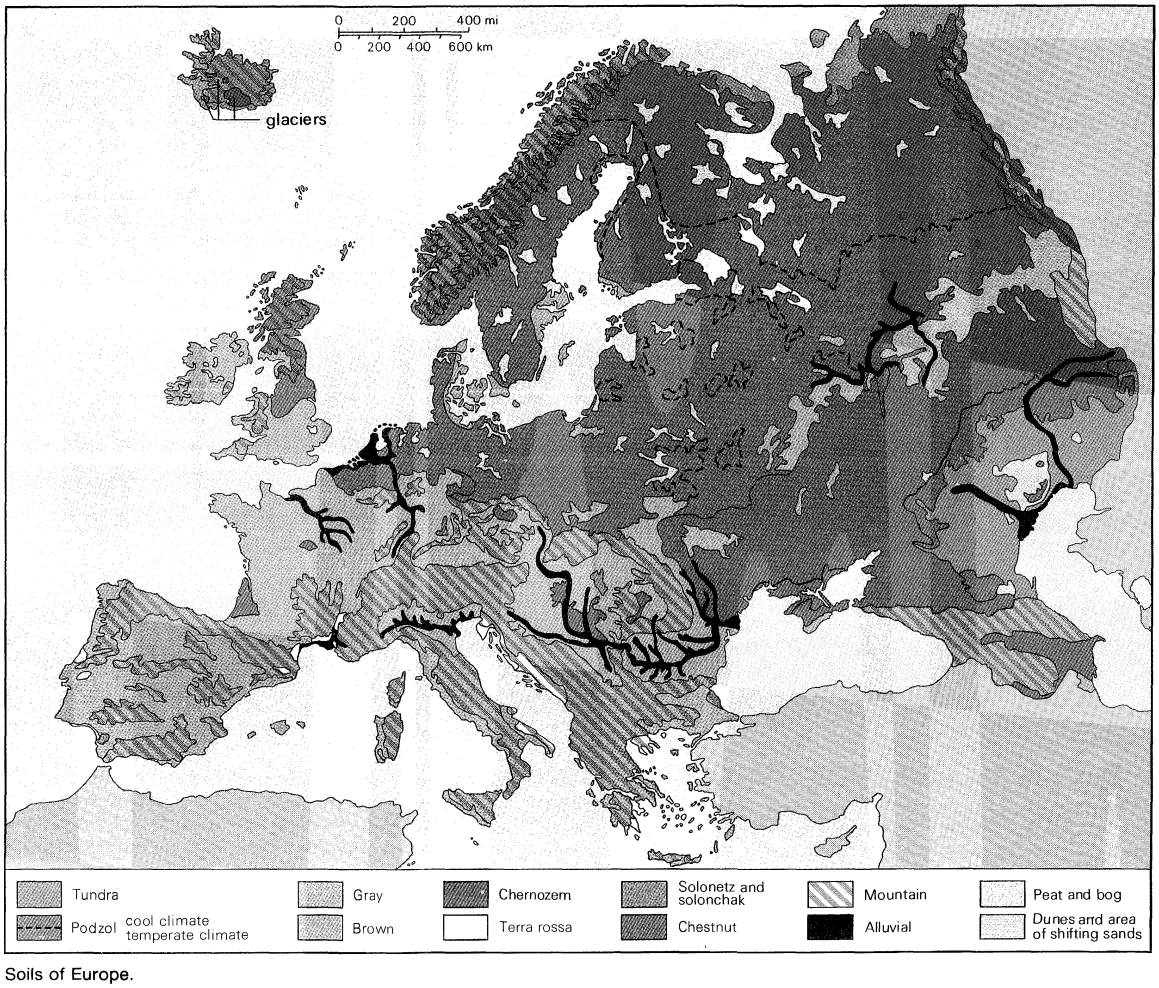
Soil fertility

Problems of classification. The origin, nature, variety, and classification of Europe's soils raise highly complex problems: so much is involved—bedrock, drainage, plant decomposition, biological action, climate, and the time factor. Humans, moreover, have done much to modify soils and, with increasing scientific knowledge, to render soils of greater and continuing value by drainage, crop rotation, and the input of suitable combinations of chemicals. In such ways, naturally poor soils can—as has been shown in Denmark—be made productive. The practice of an enforced "resting" of soils, by leaving fields fallow to recuperate, began to disappear with the agricultural revolution of the 18th century, and agronomic science continues to show how the best results can be achieved from specific soils and also how to check soil erosion. Europe's arable land lies mainly in the lowlands, which have podzols, brown, chernozem, and chestnut soils, although the upper elevation level of cultivation, as of animal husbandry, rises southward. New land is won from the sea, and this more than offsets coastal losses through erosion, but the continued losses to urban expansion and to such competitors for level land as airfields, on the other hand, have become increasingly serious.

CLIMATE

As Francis Bacon, the great English Renaissance man of letters, aptly observed, "Every wind has its weather." It is air-mass circulation that provides the main key to Europe's climate, the more so since masses of Atlantic Ocean origin can pass freely through the lowlands, except in the case of the Caledonian mountains of Norway. Polar air masses derived from areas close to Iceland and tropical masses from the Azores bring, respectively, very different conditions of temperature and humidity and produce different climatic effects as they move eastward. Continental air masses from eastern Europe have equally easy access westward. The almost continuous belt of mountains trending west–east across Europe also impedes the interchange of tropical and polar air masses.

Air-pressure belts. Patterns of some permanence controlling air-mass circulation are created by belts of air pressure over five areas. They are: the Icelandic Low, over the North Atlantic; the Azores High, a high-pressure ridge; the (winter) Mediterranean Low; the Russian High, centred over Central Asia in winter but extending westward; and the Monsoon Low, a low-pressure system over southwestern Asia. Given these pressure conditions, westerly winds prevail in northwestern Europe during the year, becoming especially strong in winter. The winter westerlies, often from the southwest, bring in warm tropical air; in summer, by contrast, they veer to the northwest



Soils of Europe.

Westerlies and the Russian High

and bring in cooler Arctic or subarctic air. In Mediterranean Europe the rain-bearing westerlies chiefly affect the western areas, but only in winter. In winter the eastern Mediterranean Basin experiences bitter easterly and northeasterly winds derived from the Russian High, and their occasional projection westward explains unusually cold winters in western and central Europe, the exceptionally warm winters of which, on the other hand, result from the sustained flow of tropical maritime air masses. In summer the Azores High moves 5°–10° of latitude northward and extends farther eastward, preventing the entry of cyclonic storms into the resultantly dry Mediterranean region. The eastern basin, however, experiences the hot and dry north and northeast summer winds called etesian by the ancient Greeks. In summer, too, the Russian High gives place to a low-pressure system extending westward, so that westerly air masses can penetrate deeply through the continent, making summer a wet season.

It is because of the interplay of so many different air masses that Europe experiences very changeable weather. Winters get sharply colder eastward, but summer temperatures relate fairly closely to latitude. Northwestern Europe, including Iceland, enjoys some amelioration due to warm Gulf Stream waters, which keep the Soviet port of Murmansk open all the year.

Climatic regions. Four regional European climatic types can be loosely distinguished, each characterized by much local topographically related variation. Further, the great cities of Europe, because of the scale and grouping of their buildings, their industrial activities, and the layout of their roads, create distinct local climates—including a central "heat island" and pollution problems.

Maritime climate. Characterizing western areas heavily exposed to Atlantic air masses, the maritime type of climate—given the latitudinal stretch of these lands—exhibits sharp temperature ranges. Thus, the January and July annual averages of Reykjavík (Iceland) and Coruña (Spain) are, respectively, 32° F (0° C) and 53° F (12° C), and 50° F (10° C) and 64° F (18° C). Precipitation is always adequate—indeed, abundant on high ground—falling the year round. The greatest amount of precipitation occurs in autumn or early winter. Summers range from warm to hot depending on the latitude and altitude, and the weather is everywhere changeable. The maritime climate extends across Svalbard, Iceland, the Faeroes, Great Britain and Ireland, Norway, southern Sweden, western France, the Low Countries, northern Germany, and northwestern Spain.

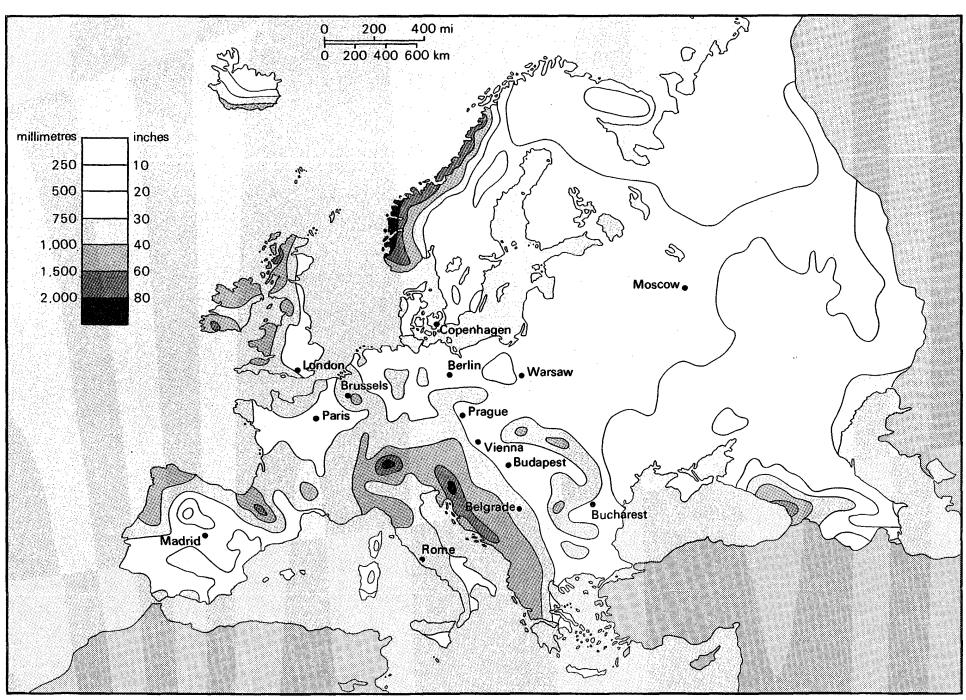
Interaction of air masses

Central European (transitional) climate. The central European, or transitional, type of climate results from the interaction of both maritime and continental air masses and is found at the core of Europe, south and east of the maritime type, west of the much larger continental type, and north of the Mediterranean type. This rugged region has colder winters, with substantial mountain snowfalls, and warmer summers, especially in the lowlands. Precipitation is adequate to abundant, with a summer maximum. The region embraces central Sweden, southern Finland, the Oslo Basin of Norway, eastern France, southwestern Germany, and much of central and southeastern Europe. The range between winter and summer temperatures increases eastward, while the rainfall can exceed 80 inches (2,000 millimetres) in the mountains, with snow often lying permanently around high peaks. The Danubian region has only modest rainfall (24 inches per year at Budapest), but the Dinaric Alps experience heavy cyclonic winter, as well as summer, rain.

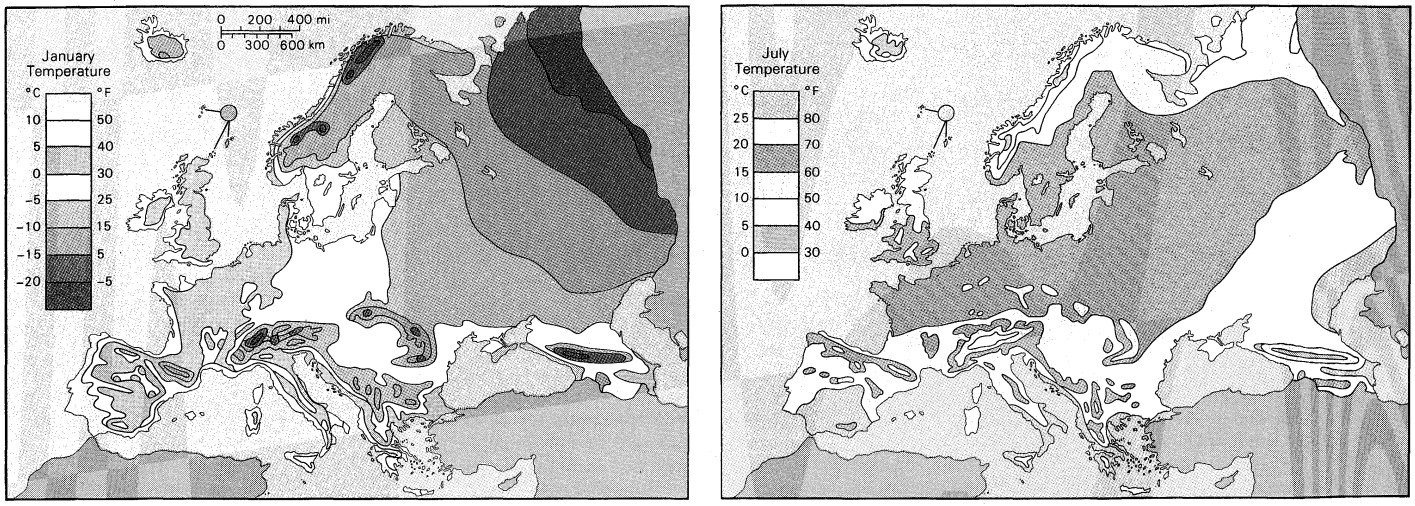
Continental climate. The continental type of climate dominates a giant share of Europe, covering the European Soviet Union (except the Baltic areas), most of Finland, and northern Sweden. Winters—much colder and longer, with greater snow cover than in western Europe—are coldest in the northeast, and summers are hottest in the southeast; the January to July mean temperatures range from 50° to 70° F (10° to 21° C). Summer is the period of maximum rain, which is less abundant than in the west: Moscow's annual average is 25 inches, while, in both the north and southeast of the Russian Plain, precipitation reaches only between 10 and 20 inches annually. In parts of the south, the unreliability of rainfall combines with its relative scarcity to raise a serious aridity problem.

Mediterranean climate. The subtropical Mediterranean climate characterizes the coastlands of southern Europe, being modified inland (for example in the Meseta Central, the Apennines, and the North Italian Plain) in response to altitude and aspect. The main features of this climatic region are mild and wet winters, hot and dry summers, and clear skies, but marked regional variations occur between the lands of the western and the more southerly lying eastern basins of the Mediterranean; the former are affected more strongly by maritime-air-mass intrusions. Rainfall in southern Europe is markedly reduced in areas lying in the lee of rain-bearing westerlies: Rome has an annual mean of 26 inches, but Athens has only 16 inches.

The effects of climate. The local and regional effects of climate on the weathering, erosion, and transport of rocks



Average annual precipitation for Europe.



Average temperatures for January (left) and July (right) in Europe.

Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.

clearly contribute much to the European landscape, and the length and warmth of the growing season, the amount and seasonal range of rainfall, and the incidence of frost affect the distribution of vegetation. Wild vegetation in its turn provides different habitats for animal life. Climate is also an important factor in the making of soils, while modern European industry and urban life depend increasingly on water supplies, with rivers and lakes continuing to provide important commercial waterways in some areas. The winter freeze in northern and eastern Europe is another effect of climate, and the spring thaw, by creating floods, impedes transport and harasses farmers. The snow cover of the more continental regions is useful to people, however, for it stores water for the fields and provides snow for sled users.

Regional variations of climate also help determine where crops are grown commercially. In southern Europe the climate supports specially adapted wild vegetation and precludes all-year grass in coastal lowlands, while the practice of moving flocks and herds to pastures seasonally available at different altitudes is clearly adapted to other conditions set by climate. In sum, in only a modest proportion of Europe does climate somewhat restrict human occupation and land use. These areas include regions of high altitude and relief, such as the subarctic highlands of the Scandinavian Peninsula and Iceland, the Arctic areas along the White Sea of the northern Soviet Union, and the arid areas of interior Spain.

PLANT LIFE

Major vegetation zones. The terms "natural," "original," and "primitive," as epithets applied to the vegetation of Europe, have no precise meaning unless they are related to a specific time in geologic history. It is, nevertheless, possible to envisage continental vegetation zones as they formed and acquired some stability during postglacial times, although such zones are only rarely recalled by present-day remnants.

The tundra. Tundra vegetation, made up of lichens and mosses, occupies a relatively narrow zone in Iceland and the extreme north of the European Soviet Union and Scandinavia, although this zone is continued southward in the mountains of Norway. Vegetation of a similar kind occurs at altitudes of 5,000–6,000 feet in the Alps and the northern Urals.

The boreal forest. Southward, the virtually treeless tundra merges into the boreal (northern) forest. The more northerly zone is "open," with stands of conifers and with willows and birch thickets rising above a lichen carpet. It is most extensive in the northern European Soviet Union but continues, narrowing westward, across Sweden. South of this zone, and with no abrupt transition, the "closed" boreal forest occupies a large fraction—mainly north of the upper Volga—of the Soviet Union proper and Scandinavia. Conifers, thin-leaved and resistant to cold, together with the birch and larch, predominate.

The mixed forest. The northern vegetation may superficially suggest its primeval character, but the zone of mixed forest that once stretched across the continent from Great Britain and Ireland to the central Soviet Union has been changed extensively by humans. Only surviving patches of woodland—associations of summer-leaf trees and some conifers, summarily described as Atlantic, central, and eastern—hint at the formerly extensive cover.

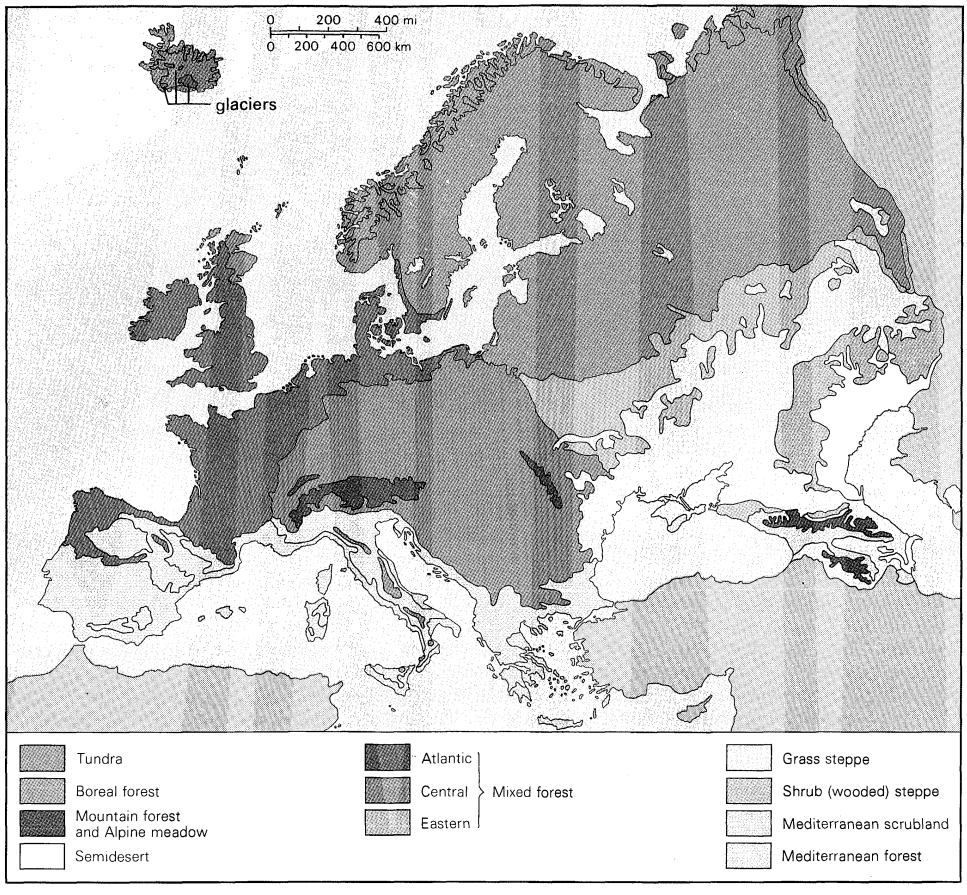
The Mediterranean complex. In southern Europe, Mediterranean vegetation has a distinctive character, containing hard-leaf forests and secondary areas of scrub, especially maquis (macchie), which is made up of trees, shrubs, and aromatic plants. Such scrub is scattered because of summer drought, particularly in areas where the soil is underlain by limestone or where there is little, if any, soil.

Steppe and semidesert. Three further vegetation zones are confined to the southeastern European Soviet Union, although the first two—the wooded steppe and the grass steppe—extend into the Danubian lowlands. Finally, semidesert vegetation characterizes the dry lowland around the northern shores of the Caspian Sea.

The shaping of vegetation zones. *Climatic change.* The primeval vegetation of Europe began to take shape as the climate ameliorated following the retreat of the Pleistocene ice sheets. The microscopic study of pollen grains preserved in datable layers of peat and sediments has made it possible to trace the continental spread, in response to

H. Fristedt—Carl E. Ostman ab

Coniferous forests and lakes on the ancient Baltic Shield of Finland.



Vegetation zones of Europe.

The Alpine barrier

climatic improvement, of forest-forming trees. The double barrier of the Alps and the Mediterranean Sea had checked the retreat of trees at the onset of the Great Ice Age, and there were relatively few indigenous species to return northward from unglaciated refuges. In the first postglacial climatic phase (the boreal), spruce, fir, pine, birch, and hazel nevertheless established themselves as far north as central Sweden and Finland. During the succeeding climatic optimum (the Atlantic phase), which was probably wetter and certainly somewhat warmer, mixed forests of oak, elm, common lime (linden), and elder spread northward. Only in the late Atlantic period did the beech and hornbeam spread into western and central Europe from the southeast.

During postglacial times, therefore, when small numbers of humans were living within Europe, the continental surface was thickly clad with trees and undergrowth, except where tree growth was precluded by extreme cold, high altitude, bad drainage, or exposure to persistent gales. Even those relatively attractive areas where windblown loess had been deeply deposited are now known to have had woods of beech, hawthorn, juniper, box, and ash, as did also limestone plateaus. The Mediterranean peninsulas also had evergreen and mixed forests rooted in an ample soil.

The role of humans. From prehistoric times onward, with ever-increasing force, humans, seeking optimum economic use of available resources, have acted as a vigorous agent of vegetation change. The effects of grazing animals may well explain why some heathlands (*e.g.*, the Lüneburg Heath in northeastern West Germany) replaced primeval forest. By fire and later by ax, forest clearance met demands for homes and ships, for fuel, for charcoal for iron smelting, and, not least, for more cultivation and pasture. The mixed boreal forests suffered most because their relatively rich soils and long and warm growing season promised good returns from cultivation. The destruction of woodlands was markedly strong when population was growing (as between about AD 800 and 1300). It was later intensified by German colonization east of the Rhine and reached maximum scale in the 19th century. In southern Europe—where naval demands were continuous and sources of suitable timber sharply localized—tree cutting entailed, from classical antiquity onward, serious soil loss through erosion, increased aridity, floods, and marsh formation. Farther north throughout the continent, as present distribution of arable land shows, former forests were reduced to remnants; only in the north and below the snow line of Alpine mountains have forests of large and continuing commercial value survived. These coniferous forests of Sweden, Finland, and the northern Soviet Union are "cropped" annually to preserve their capital value. On the more positive side may be noted the reclamation of marshlands and the soil improvement of hill grasslands and heaths, their wild vegetation being replaced by pasture and crops; in timber-deficient countries the afforestation of hillslopes, chiefly with quickly growing conifers, belatedly attempts to restore some of the former forests. Another drastic vegetation change brought about by humans has been the virtual elimination of the wooded and grass steppes, which have become vast granaries.

Afforestation in Scandinavia and the Soviet Union

Exterior influences and European survivals. To a surprising degree, European vegetation stemmed from the importation of plants from other continents, although some imported crops—notably citrus fruits, sugarcane, and rice—can only grow marginally in Europe, and then by irrigation. From an original home of wild grasses in Ethiopia, cultivated varieties of wheat and barley reached Europe early, via the Middle East and Egypt, as did also the olive, the vine, figs, flax, and some varieties of vegetables. Rice, sugarcane, and cotton, of tropical Indian origin, were introduced by the Arabs and Moors, especially into Spain. The citrus fruits, peaches, mulberries, oats, and millet reached Europe from original Chinese habitats, and Europe owes corn (maize), tobacco, squashes, tomatoes, red peppers, prickly pears, agave (sisal), and the potato—first grown for fodder but destined to become the cheap staple food for the large families of low-paid workers of the 19th century—to the Americas. Europe has drawn greatly

Deciduous forest of beech in autumn, New Forest, southern England, U.K.
Heather Angel—Biofotos

on East Asia and North America for trees, especially ornamental trees, while some acacias and the eucalyptus derive from Australia. The sugar beet, however, was a European discovery, first grown when much of Napoleonic Europe was subjected to maritime blockade.

The forests of northern Europe and the Alpine ranges, although in no sense primeval, represent unchanging land use during the postglacial period. The "closed boreal forest" occupies some one million square miles (2.6 million square kilometres), made up of a spruce–fir association (but with stands of pine, birch, and larch) above an undergrowth of mosses and herbs. This large and valuable reserve of timber is of world importance; forests once covered 80 percent of Europe's surface, and they still occupy about 30 percent.

Human adaptations. Clearly, animal life, wild and domesticated, has been adjusted to fit largely man-made patterns of vegetation, which, in turn, reflect agelong attempts to achieve chiefly economic ends. With such endeavours are associated varieties of *modes de vie,* or "modes of livelihood." In the mountains as in the boreal forests, the environment is exploited by winter lumbering and by the transport of felled trees by river after the spring thaw. So, too, agriculture in its many forms—in part for subsistence but commonly for urban markets—is a basic occupation of the lowlands, long cleared of extensive forests or steppe vegetation. In Mediterranean Europe, rural life, based on horticulture and arboriculture rather than on large-scale cultivation, as well as on the rearing of sheep and goats and wheat cultivation, continues, little changed in many areas. For such deeply rooted fruit-bearing trees as the olive and vine, use is made of sloping, broken, and terraced land. Farming also extends to specialized forms with respect to the subtropical crops that climate, sometimes supplemented by irrigation, permits.

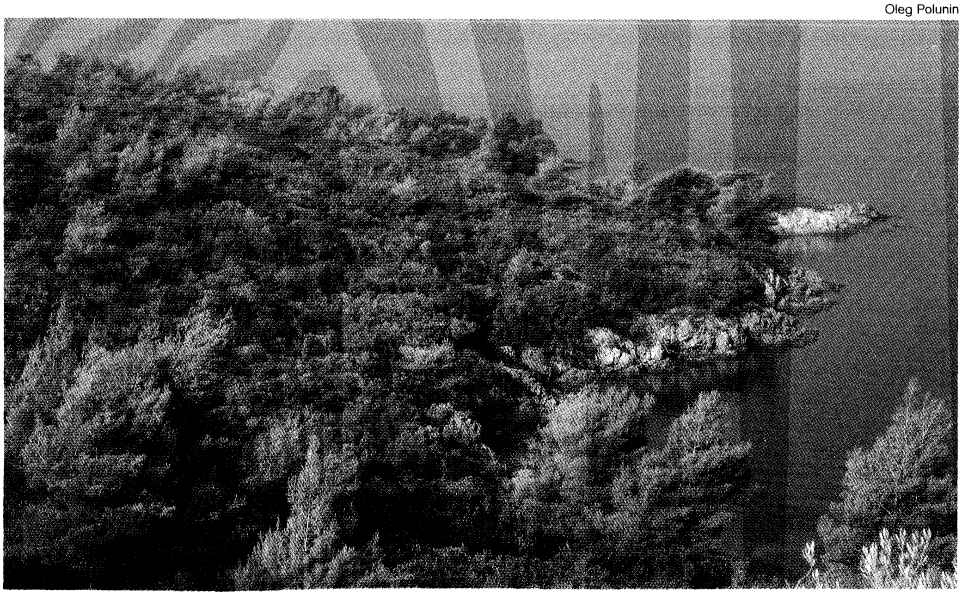
Modes de vie

ANIMAL LIFE

Patterns of distribution. With animals as with plants, the earlier Pleistocene range and variety has been much reduced since man disposed of what nature provided. Wild fauna has been long in retreat since Upper Paleolithic times, when, as cave drawings portray, small human groups held their own against such big game as aurochs and mammoths, now extinct, and also against such survivors as the elephant, bison, horse, and boar. Hares, swans, and geese were also hunted, and salmon, trout, and pike were fished. Humans were, inevitably, the successful competitor for land use. By prolonged effort settlers won the land for crops and for domesticated animals, and they hunted animals, especially for furs. As population mounted in industrializing Europe, humans no less inevitably destroyed, or changed drastically, the wild vegetation cover and the animal life. With difficulty, and largely on human sufferance, animals have nevertheless survived in association with contemporary vegetation zones.

The tundra. In the tundra some reindeer (caribou), both wild and domesticated, are well equipped to withstand the cold. Their spoon-shaped hooves are useful in finding food in rough ground. Their herds migrate southward in winter and eat lichens and other plants, as well as flesh, notably that of lemmings and voles. Dogs, too, are reared for traction but yield less than reindeer, which also provide meat, milk, pelts, wool, and bone. The Arctic fox, bear, ermine, partridge, and snowy owl may appear in the

Oleg Polunin



Maquis (macchie) vegetation on the Mediterranean coast, near Sithonía, Greece.

Steppe grasslands at Point Kaliakra, Bulg., on the northwestern shore of the Black Sea.
Oleg Polunin

tundra, where, in the short summer, seabirds, river fish, and immigrant birds (swans, ducks, and snipes) vitalize a harsh environment then made almost intolerable by the swarms of biting midges.

The thinning out of animal species in boreal forests

Boreal associations. In the boreal forests the richness of animal and bird life, which had persisted throughout historical times, has now been greatly reduced. Among large surviving ungulates are the elk (moose), reindeer, and roebuck, and, among big beasts of prey, the large brown bear. The lynx has been exterminated by humans, but not the wolf, fox, marten, badger, polecat, and white weasel. The sable, which is much hunted for its fur, only just survives in the northeastern forests of the European Soviet Union. Rodents in the forests include squirrels, the white Arctic hare, and (in the mixed forests) the gray hare and the beaver. Among birds are the black game, snipe, hazel hen, white partridge, woodpecker, and crossbill, all of which assume protective colouring and are specially adapted to be able to find their food in a woodland environment. Owls, blackbirds, tomtits, and bullfinches may be seen in the forests, and, in meadow areas, geese, ducks, and lapwings may be seen.

The steppes. The fauna of the steppe zones now lacks large animals, and the saiga antelope has disappeared. Numerous rodents, including the marmot, jerboa, hamster, and field mouse, have increased in numbers to become pests, now that nearly all the steppe is under cultivation. Equally plentiful birds include the bustard—who can fly as well as run—quail, gray partridge, and lark. These take on yellowish gray or brown protective colouring to match the dried-up grass. Eagles, falcons, hawks, and kites comprise the birds of prey; water and marsh birds—especially the crane, bittern, and heron—also make their homes in the steppes. Different kinds of grasshopper (locusts) and beetles are insect pests.

Mediterranean and semidesert associations. In Mediterranean Europe, remnants of mountain woodland harbour wild goats, wild sheep—such as the small mouflon of Corsica and Sardinia—the wildcat, and wild boar. Snakes, including vipers, and lizards and turtles are familiar reptiles, but birds are few. The faunas of the semidesert areas of the southeastern European Soviet Union also show affiliations with the grass steppe and the desert between which they lie. Two types of antelope (saiga and jaran) survive there, as do rodent sand marmots and desert jerboa and, as a beast of prey, the sand badger. There are many reptiles—lizards, snakes (cobras and steppe boas)—and tortoises. The saksaul jay and the saksaul sparrow, named after the desert tree, also live there, while scorpions, the karakurt spiders, and the palangid are insects dangerous to humans and camels.

Conservation problems. Pressure on space, hunting, either for sport or to protect crops, the pollution of seawaters and fresh waters, and the contamination of cropland have so reduced many animal species that strong efforts are now being made to preserve those threatened with extinction, in such refuges where they still, precariously, live.

Nature reserves

Nature reserves have been set up in many European countries, with international support from the International Union for Conservation of Nature and Natural Resources and the World Wildlife Fund. Seabirds find safe homes, for example, in the Lofoten Islands of Norway and the Farne Islands of northeastern England. The snowy owl, which feeds on lemmings, is seen in Lapland, the rare great bustard in the Austrian Burgenland, and the musk-ox in Svalbard. Père David's deer, which had become extinct in China, its native home, was introduced in 1898 at Woburn Abbey, Eng., where it now flourishes. Nearly half the bird species of Europe, including the egret and the imperial eagle, are represented in the Doñana National Park, within a setting of wild vegetation in the Las Marismas region of the Guadalquivir estuary in southwestern Spain; there too the Spanish lynx survives. In Poland the extensive Białowieza National Park, a wild forest once hunted yearly by the tsars, contains deer, wild boar, elk (moose), bears, lynx, wolves, eagle owls, black storks, the European bison (wisent), and the tarpan, a gray-coloured horse and a survivor from remote days. Contiguous with the forest, in Belorussia, is the Belovezhskaya Pushcha, where European bison are preserved. Italy has its famous Gran Paradiso National Park in the Aosta Valley, which preserved from extinction the Alpine ibex; Austria has a bird refuge in Neusiedler Lake, which is the only breeding site of white egrets in western Europe; and the huge and magnificent delta of the Danube is largely left to wildlife. The golden eagle, Alpine marmots, and chamois are to be seen in the German Alps near Berchtesgaden.

Other rare birds are the sandwich tern, at Norderoog Island, W.Ger., and the spoonbill and cormorant, found, respectively, at Texel Island and Norderoog Island, the former in The Netherlands. For ornithologists (as for botanists) Iceland has abounding interest, notably at Slúttnes, an island in the shallow Lake Mývatn. The beautiful wild horses of Camargue Nature Reserve (Rhône delta), the wild ponies of the New Forest (England), and the Barbary apes, maintaining a politically disinterested foothold on the Rock of Gibraltar, continue undiminished in popular interest.

Thus, the European environment, once not so unequally shared by plants, animals, and people, has, with the march of civilization, been subjected to the attempt at mastery by humans. Favoured by their proximity to the Middle East, where crop cultivation and animal domestication first began, Europeans have fashioned cultural landscapes, at the expense of wild nature, to serve their economic and social ends. Only with difficulty—and sporadically—has wild nature survived, and only just in time has awareness of the cultural losses from the impoverishment of natural vegetation and its animal associates underlined the urgent need for careful protection and preservation of nonhuman nature for communal enjoyment and scientific research.

Against certain pests, notably the anopheles mosquito and the rabbit, war has been waged with good effect, for malaria no longer afflicts Mediterranean Europe, and rabbits, competitors for grass, have been greatly reduced. On the credit side, too, should be listed the full use made of domesticated animals for pastoral husbandry—on high and rough ground, as well as on farms. The familiar farm animals are selectively bred and reared with some regard to the physical character of their environment as also to market demands and government decisions. In the far north, herds of reindeer are adapted to withstand cold and to find their food in snow-covered ground. In the rough, hilly, scrubland of Mediterranean Europe, the sheep, goat, donkey, mule, and ass have adapted well. The horse, which, in its long history, has drawn chariots, carried mounted knights, and hauled the plow, wagons, artillery, stagecoaches, canalboats, and urban trams, is now largely replaced by the tractor, truck, and jeep. The horse is now raised more for racing, riding, ceremonials, and the hunting of fox and stag but is still used for farm work, especially in eastern Europe. Distribution maps of animals kept on farms show how widely they enter into farming: sheep have a special concentration in Great Britain and the Balkan countries, cattle have a small place in southern Europe, while pigs are relatively numerous in the north, especially in the highly populated areas of West Germany, Denmark, and the Low Countries.

The people

The vast majority of Europe's inhabitants are of the European (or Caucasoid) geographic race, characterized by white or lightly pigmented skins and variability in eye and hair colour and by a number of biochemical similarities; there is also an increasing number of people of African and Asian ancestry, although their proportion of the population is still small.

Origins of the Europeans

The origins of the Europeans as a distinct group may never be learned. It is known, however, that the continent had a scanty population of now-extinct hominid species before modern humans appeared some 40,000 years ago and that throughout its prehistoric period it received continual waves of immigrants from Asia. The legacy of these immigrants can be seen in the variety of physical types and cultural features that are found throughout Europe.

CULTURAL PATTERNS

Culture groups. Efforts have been made to characterize different "ethnic types" among European peoples, but these are merely selectively defined physical traits that, at best, have only a certain descriptive and statistical value. On the other hand, territorial differences in language and culture are well known; these have been of immense social and political import in Europe.

These differences stand in sharp contrast to such relatively recently settled lands as the United States, Canada, and Australia. Given the agelong occupation of its soil and minimal mobility for the peasantry—long the bulk of the population—Europe became the home of many linguistic and national "core areas," separated by mountains, forests, and marshlands. Its many states, some long-established, introduce another divisive element that was augmented by modern nationalistic sentiments. Efforts to associate groups of states for specific defense and trade functions appear, slowly, to point toward wider unitary associations, with fundamental East–West differences. Europe thus presents two clear-cut, opposing units, open to trade and cultural exchanges, and a number of such relatively neutral states as Ireland, Sweden, Austria, Switzerland, Yugoslavia, Finland, and Spain.

"Core areas"

The map showing the distribution of European ethnic culture areas identifies a total of about 160 different groups. Of these, about three-fifths are found principally west of the Soviet Union, and about two-fifths are situated mainly within Soviet Europe or the Transcaucasus region. Each of these large groups exhibits two significant features. First, each is characterized by a degree of self-recognition by its members, although the basis for such collective identity varies from group to group. Second, each group—except the Jews and Gypsies—tends to be concentrated and numerically dominant within a distinctive territorial homeland.

For a majority of groups the basis for collective identity is possession of a distinctive language or dialect. The Catalans and Galicians of Spain, for example, have languages notably different from the Castilian of the majority of Spaniards. On the other hand, some peoples may share a common language yet set each other apart because of differences in religion. In Yugoslavia, for instance, the Eastern Orthodox Serbs, Muslim Bosnians, and Roman Catholic Croats all speak the Serbo-Croatian language, but the members of each group generally have antagonistic views toward the others. Some groups may share a common language but remain separate from each other because of differing historic paths. Thus, the Walloons of southern Belgium and the Jurassians of Switzerland both speak French, yet they see themselves as quite different from the French because their groups have developed almost completely outside the boundaries of France. Even when coexisting within the same state, some groups may have similar languages and common religions but remain distinctive from each other because of separate past associations. In Czechoslovakia the historic linkages of Slovaks with the Hungarian kingdom and Czechs with the Austrian state have kept the two groups apart despite more than seven decades of coexistence within a single country.

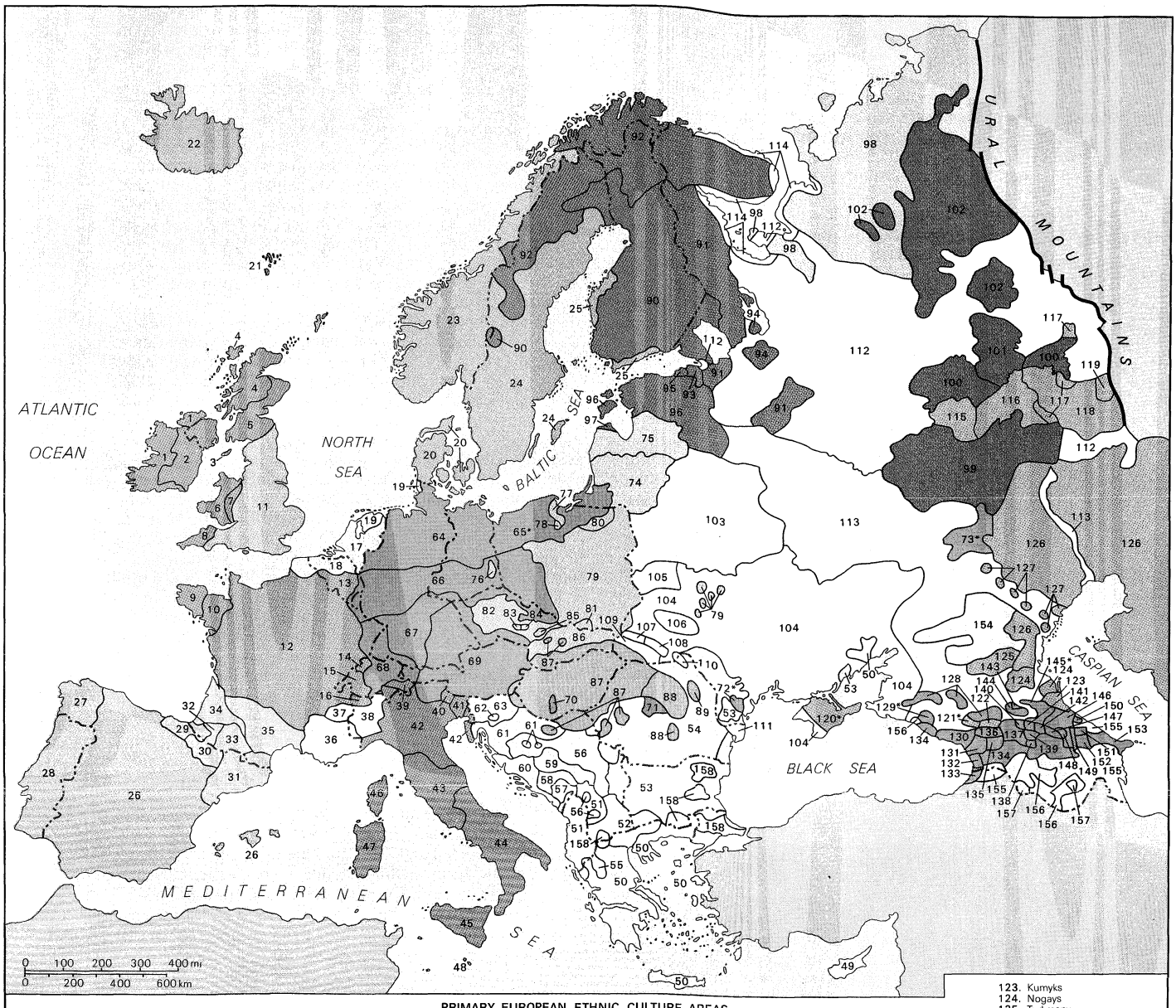
The primary groups of the map have been associated by ethnographers into some 21 culture areas, 15 of which lie west of the Soviet border. The grouping is based primarily on similarities of language and territorial proximity. Although individuals within a primary group generally are aware of their cultural bonds, the various groups within an ethnographic culture area do not necessarily share any self-recognition of their affinities to one another. This is particularly true in the Balkan culture area. Peoples in the Scandinavian and German culture areas, by contrast, are much more aware of belonging to broader regional civilizations.

Languages. *Romance, Germanic, and Slavic languages.* Within the complex of European languages, three major divisions stand out: Romance, Germanic, and Slavic. All three are derived from a parent Indo-European language of the early migrants coming to Europe from southwestern Asia.

The Romance languages dominate western and Mediterranean Europe and include French, Spanish, Portuguese, Italian, and Romanian, plus such lesser-known languages as Occitan (Provençal) in southern France, Catalan in northeastern Spain and Andorra, and Romansh in southern Switzerland. All are derived from the Latin language of the Roman Empire.

The Germanic languages are found in central, northern, and northwestern Europe. They are derived from a common tribal language that originated in southern Scandinavia and Denmark, and they include German, Dutch, Danish, Norwegian, Swedish, and Icelandic, as well as the minor Germanic tongues of Frisian in the northern Netherlands and northwestern West Germany and Flemish in the Flanders provinces of northern Belgium and adjacent parts of northern France. English is a Romance-Germanic hybrid with extensive vocabularies from both linguistic strains.

The Slavic languages are characteristic of eastern and southeastern Europe, including the Soviet Union. These languages are usually divided into three branches: West,



PRIMARY EUROPEAN ETHNIC CULTURE AREAS

I. The Atlantic fringe

Celtic culture area
1. Gaelic-speaking Irish
2. English-speaking Irish
3. Manx
4. Gaelic-speaking Scottish
5. English-speaking Scottish
6. Celtic-speaking Welsh
7. English-speaking Welsh
8. Cornish
9. Celtic-speaking Bretons
10. French-speaking Bretons

English culture area
11. English

French culture area
12. French
13. Walloons
14. Jurassian Swiss
15. Franco-Swiss
16. Valdotans

Netherlandish culture area
17. Dutch
18. Flemings
19. Frisians

Scandinavian culture area
20. Danes
21. Faeroese
22. Icelanders
23. Norwegians
24. Swedes
25. Finnish Swedes

II. The Mediterranean

Iberian culture area
26. Spaniards
27. Galicians
28. Portuguese

Pyrenean culture area
29. Basques
30. North Aragonese
31. Catalans
32. Béarnese
33. Bigordans
34. Gascons
35. Lower Languedocians

Mediterraneo-Alpine culture area
36. Provençals
37. Savoyards
38. Piedmontese

Italic culture area
39. Rhaeto-Romansh
40. Ladins
41. Friulians
42. North Italians
43. Middle Italians
44. South Italians
45. Sicilians
46. Corsicans
47. Sardinians
48. Maltese

Balkan culture area
49. Cypriots
50. Greeks
51. Albanians
52. Macedonians
53. Bulgarians
54. Romanians
55. Aromuns
56. Serbs
57. Montenegrins
58. Hercegovinians
59. Bosnians
60. Dalmatians
61. Croats
62. Chich
63. Slovenes

III. The Carpatho-Alpine-West Plain Climax

German culture area
64. West Lower Germans
65. East Lower Germans*
66. Middle Germans
67. Upper Germans
68. Alemanni
69. Bayuvars
70. Transdanubian Swabians
71. Transylvanian Saxons
72. Moldavian Germans*
73. Volga Germans*

Baltic culture area
74. Lithuanians
75. Latvians

West Slavic culture area
76. Sorbs (Wends)
77. Kashubians
78. Slovincz
79. Poles
80. Mazurians
81. Górale
82. Czechs
83. Horaks
84. Zahoraks
85. Hannaks
86. Slovaks

West Ugric culture area
87. Hungarians (Magyars)
88. Siculi
89. Csangos

IV. The East Plain and the Caucasus

West Finno-Permian culture area
90. Finns
91. Karelians
92. Lapps
93. Ingrians
94. Veps
95. Votes
96. Estonians
97. Livonians

Samoyed culture area
98. Nenets

East Finno-Permian culture area
99. Mordvins
100. Maris (Cheremis)
101. Udmurts (Votyaks)
102. Komis

East Slavic culture area
103. Belorussians
104. Ukrainians
105. Volhynians
106. Podolians
107. Lemks
108. Boyks
109. Ruthenians
110. Hutsuls
111. Lipovans
112. North Great Russians
113. South Great Russians
114. Pomors

Pre-Uralian Turko-Tatar culture area
115. Chuvash
116. Kazan Tatars
117. Tepters
118. Bashkirs
119. Misars

Ponto-Caspian Turko-Tatar culture area
120. Crimean Tatars*
121. Karachays*
122. Taulu
123. Kumyks
124. Nogays
125. Turkmeny
126. West Kazakh-Kirgiz
127. Lower Volga Tatars

Caucasian culture area
128. Circassians
129. Balkars*
130. Abkhaz
131. Mingrelians
132. Gurians
133. Laz
134. Georgians
135. Ymer
136. Svan
137. Ossetes
138. Ingushes
139. Pshav
140. Mtiul
141. Khevsur
142. Thusk
143. Lezgians
144. Kabardians
145. Chechens (Shishans)*
146. Avars
147. Dido
148. Andi
149. Lakks
150. Tabasarans
151. Dargins
152. Kyrinans
153. Tats

Sporadic ethnic units of extra-European origins
154. Kalmyks
155. Azerbaijani Tatars
156. Armenians
157. Kurds
158. Osmano-Turks
Gypsies
Jews

*Peoples partly dispersed by the Soviet government since 1944 (Volga Germans since 1941), but preserving the right to their native homes.

Distribution of European ethnic culture areas.

East, and South. Among the West Slavic languages are Polish, Czech and Slovak, the Wendish language of East Germany, and the Kashubian language of northern Poland. The East Slavic languages are Russian, Ukrainian, and Belorussian. The South Slavic languages include Slovene, Serbo-Croatian, Macedonian, and Bulgarian.

Other languages. In addition to the three major divisions of the Indo-European languages, three minor groups are also noteworthy. Modern Greek is the mother tongue of Greece and of the Greeks in Cypress, as well as the people of other eastern Mediterranean islands. Older forms of the language were once widespread along the eastern and southern shores of the Mediterranean and in southern peninsular Italy and Sicily. The Baltic language family includes modern Latvian and Lithuanian. The Old Prussian language also belonged to the Baltic group but was supplanted by German through conquest and immigration. Europe's Gypsies speak the distinctive Romany language, which has its origins in the Indic branch of the Indo-European languages.

Two other Indo-European language divisions were formerly widespread but now are spoken only by a few groups. Celtic languages at one time dominated central and western Europe from a core in the German Rhineland. Cultural pressures from adjacent Germanic- and Romance-speaking civilizations eliminated the Celtic culture area, save for a few remnants, including the Welsh, the Gaelic-speakers of the Scottish Highlands and western Ireland, and the Celtic-speaking Bretons of the northwestern Brittany Peninsula of France. The Thraco-Illyrian branch of the Indo-European languages formerly was spoken throughout the Balkan Peninsula north of Greece. Its cultural survival is now to be found only in Albanian.

Non-Indo-European languages are also spoken on the continent. The sole example in western Europe is the Basque language of the western Pyrenees Mountains; its origins are obscure. In northeastern Europe the Finnish, Lapp, Estonian, and Hungarian languages belong to a family of the Finno-Ugric language group that has other representatives in the middle Volga River region. Turkic languages are spoken in parts of Bulgaria and Yugoslavia and in the Azerbaijan S.S.R. and adjacent regions of the Soviet Caucasus area.

Religions. Most primary culture groups in Europe have a single dominant religion, although the English, German, and Hungarian groups are noteworthy for the coexistence of Roman Catholicism and Protestantism. Like its languages, Europe's religious divisions fall into three broad variants of a common ancestor, plus distinctive faiths adhered to by smaller groups.

Christianity. The majority of Europeans adhere to one of three broad divisions of Christianity: Roman Catholicism in the west and southwest, Protestantism in the north, and Eastern Orthodoxy in the east. The divisions of Christianity are the result of historic schisms that followed its period of unity as the adopted state religion in the late stages of the Roman Empire. The first major religious split began in the 4th century, when pressure from "barbarian" tribes led to the division of the empire into western and eastern parts. The bishop of Rome became spiritual leader of the West, while the patriarch of Constantinople led the faith in the East; the final break occurred in 1054. The line adopted to divide the two parts of the empire remains very much a cultural discontinuity in the Balkan Peninsula today, separating Roman Catholic Croats, Slovenes, and Hungarians from Eastern Orthodox Montenegrins, Serbs, Bulgarians, and Romanians. The second schism occurred in the 16th century within the western branch of the religion, when Martin Luther inaugurated the Protestant Reformation. Although rebellion took place in many parts of western Europe against the central church authority vested in Rome, it was successful principally in the Germanic-speaking areas of Britain, northern Germany, and Scandinavia, the latter including the adjacent regions of Finland, Estonia, and Latvia.

Judaism and Islām. Judaism has been practiced in Europe since Roman times. Jews undertook continued migrations into and throughout Europe, in the process dividing into two distinct branches, the Ashkenazi and Sephardi. Although through persecution and emigration their numbers are much reduced in Europe—particularly in eastern Europe, where Jews once made up a large minority population—Jews are still found in urban areas throughout the continent.

Islām also has a long history in Europe. Islāmic incursions into the Iberian and Balkan peninsulas have been influential in the cultures of those regions. Muslim communities still exist in several parts of the Balkans, including European Turkey, Albania, central Yugoslavia, and northeastern Bulgaria. Muslims are more numerous in the European Soviet Union, including the Kazan Tatars and Bashkirs in the Volga-Ural region and the Azerbaijani and lesser groups in the Caucasus region.

DEMOGRAPHIC PATTERNS

Europe has always been one of the most populous parts of the world. Although its estimated population numbered only one-third of Asia's in 1650, 1700, and 1800, this nevertheless accounted for one-fifth of humanity. Despite large-scale emigration, this proportion increased to one-quarter by 1900, when Europe's total just exceeded 400 million. Such high numbers, achieved by high birth rates and falling death rates, were sustained by expanding economies. As numbers have grown proportionately faster in the Americas, Asia, and Africa, Europe's population has fallen to less than one-seventh of the world total.

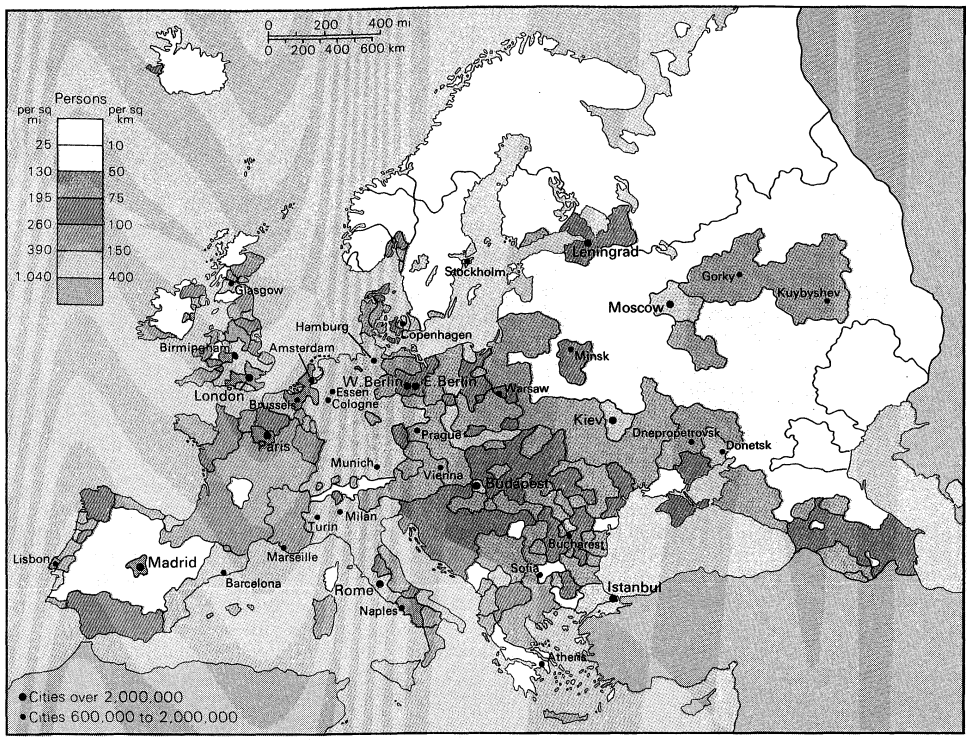
Overall densities. In antiquity the focus of settlement was in southern Europe; but the south lost its numerical domination as, from medieval times onward, settlement developed vigorously in western and central Europe and as, later still, the steppelands of the Ukraine and Hungary were settled for crop farming. While northern Europe, from Iceland and the Scottish Highlands to the northern European Soviet Union, is only scantily settled, the population reaches high densities in a more southerly belt, stretching from England across northern France and industrial Germany to the Moscow region. A second major population strip extends southward from the Ruhr valley through Italy. High populations are often associated with coalfields that, in the past more than today, strongly attracted industry, although giant cities like London, Paris, and Leningrad, offering large markets and labour forces, have created regions of high density. Other populous areas are sustained by mining, industry, commerce, and productive agriculture. The Netherlands is the most densely populated country; Iceland and Norway are the least dense. Population is scantiest in mountain regions, some highlands, arid parts of Spain, and the Arctic regions of the Soviet Union.

High-density areas

Urban and rural settlement. With easier travel and the lure of developing industrial areas, many culturally rich, high-altitude areas have suffered severe depopulation. Urbanization—offering varied employment, better social services, and, apparently, a fuller life—has further reduced the rural population, a drift aided by the mechanization of agriculture. City life has, from classical antiquity, nurtured European culture, although tributary rural life was for centuries the common lot. During the 19th and 20th centuries, however, there has been a revolutionary urbanization that embraces the majority of contemporary Europeans. Some towns are old, containing architectural survivals from their historic past; many more are creatures of the Industrial Revolution.

Nearly three out of four Europeans west of the Soviet border now live in cities. In most of the highly industrialized countries the proportion of urban dwellers is high: more than 90 percent in the United Kingdom, almost 90 percent in The Netherlands, and more than 80 percent in West Germany, Denmark, and Sweden. Others with figures greater than 70 percent include East Germany, Czechoslovakia, France, and Belgium. Only Portugal, Albania, and Yugoslavia have urban populations that number less than half of their national totals. Towns of different scale and varying function continue to grow rapidly, usually in concentric rings outward from the original core. Europe contains many cities of more than one million population, more than one-fifth of the world total, and many of the more highly industrialized parts of the

Conurbations



Population density of Europe.

continent are marked by giant, sprawling metropolitan areas. One distinct type is represented by the conurbation resulting from outgrowth from London; another, as in the Ruhr, by the fusing together of separate cities. Both areas stem from an unchecked industrial expansion associated with population growth—including immigration from rural areas. As elsewhere in the world, these giant agglomerations pose difficult social and aesthetic problems; but by concentrating population they help to preserve the countryside from becoming too built-up.

Population trends. Western and northern Europe took the lead in the medical and social "death controls" that since the mid-19th century have sharply reduced infant mortality and lengthened life expectancy. Although infant mortality rates remain relatively high in Yugoslavia, Romania, Hungary, and Poland, low mortality rates have been achieved virtually everywhere else in Europe.

Birth rates and death rates, as they vary in time and place, necessarily affect the proportion of manpower available to the different European countries for the economy and the armed forces. In most countries, increased longevity and lowered birth rates have generated a rising proportion of retired citizens. Also, the trend toward education over longer periods draws more young people from the economy. The labour force thus shrinks somewhat, although it is everywhere (except in Spain, Malta, Ireland, and Greece) more than 40 percent of the population, and in Denmark, Bulgaria, Finland, Sweden, the Soviet Union, and East Germany it exceeds 50 percent. In many European countries labour force totals have remained high, but this is primarily because of greater employment of women.

Emigration and immigration. Despite heavy mortality resulting from continual wars, Europe has always been in modern times a generous source of emigrants. Since the geographic discoveries of the late 15th century, both "push" and "pull" factors explain an exodus greatly accelerated by modern transportation. The push factors were often sheer poverty, the desire to escape from persecution, or loss of jobs through economic change. The pull factors included new opportunities for better living, often at the expense of original inhabitants elsewhere. All of Europe shared in this huge transfer of population, which affected the settlement and economic development of the Americas, Australia, southern Africa, and New Zealand. Through their involvement in the horrors of the African slave trade, Europeans also produced forced migrations of nonwhite peoples that were to have immense consequences in the Old and New Worlds. Since the early 19th century an estimated 60 million people left Europe for overseas; more than half settled in the United States. Ireland lost much of its population following the potato famine of the 1840s. Northwestern Europe—Great Britain, Scandinavia, and the Low Countries—contributed the largest share of emigrants, who settled, above all, where English was spoken. Emigrants from central, eastern, and southern Europe moved later, many in the early decades of the 20th century. Affinities of languages, religion, and culture clearly explain migration patterns; South American countries, for example, had more appeal to Spanish, Portuguese, and Italians. It has been estimated that emigration from 1846 to 1932 reduced the growth rate of Europe's population by three per 1,000 per annum. The year 1913 marked a peak, with at least 1.5 million—one-third Italian and more than a quarter British—migrating overseas. Subsequent entry restrictions in the United States reduced this flood. During the late 20th century, European migrants sought new homes mainly in Australia, Canada, South America, Turkey, and the United States.

Despite high population densities, some European countries still attract settlers from other continents, mainly because their expanding economies involve labour shortages. Thus France, to increase a manpower depleted by war losses and low 1930s birth rates, has received numerous French-Algerians (as well as other Europeans, including Turks) to supplement its labour force. The United Kingdom, which steadily supplies immigrants to Australia and Canada and specialist workers to the United States, has also attracted immigrants, notably Commonwealth citizens. These immigrants, who largely provide workers for the construction industries, transport, hospitals, and domestic service, include also doctors, scholars, and businesspeople. Some, having established themselves, are able to provide homes for their immigrant relatives. There is also a continued and significant migration of Russians and Ukrainians from the European to the Asian Soviet Union.

Intra-continental migrations

Within the continent there always has been some mobility of population, high during prehistoric times and well marked during the period of decline and fall of the Roman Empire in the West, when many tribal groups—especially of Germans and Slavs—settled in specific regions where they grew into distinctive nations. During and after World War II many Germans from outlying settlements

in central and east central Europe returned to western Germany, some as forced migrants. Many eastern Europeans, too, made their way to the West until the sealing of the East–West border curtailed this flow. Migrants are chiefly workers, seeking temporary work and, often with less success, new homes. The countries of the European Economic Community (EEC, or Common Market) draw workers from southern Europe, as does Switzerland. Two other conspicuous forms of mobility in Europe are the daily commuting of city workers and the increasing movements of tourists.

The economy

Europe was the first of the major world regions to develop a modern economy based on commercial agriculture and industrial development. Its successful modernization can be traced to the continent's rich endowment of economic resources, its history of innovations, the evolution of a skilled and educated labour force, and the interconnectedness of all its parts—both naturally existing and man-made—which facilitated the easy movement of massive quantities of raw materials and finished goods and the communication of ideas.

Europe's economic modernization began with a marked improvement in agricultural output in the 17th century, particularly in England. The traditional method of cultivation involved periodically allowing land to remain fallow; this gave way to continuous cropping on fields that were fertilized with manure from animals raised as food for rapidly expanding urban markets. Greater wealth was accumulated by landowners at the same time that fewer farmhands were needed to work the land. The accumulated capital and abundant cheap labour created by this revolution in agriculture fueled the development of the Industrial Revolution in the 18th century.

The revolution began in northern England in the 1730s with the development of water-driven machinery to spin and weave wool and cotton. By mid-century James Watt had developed a practical steam engine that emancipated machinery from sites adjacent to waterfalls and rapids. Britain had been practically deforested by this time, and the incessant demand for more fuel to run the engines led to the exploitation of coal as a major industry. Industries were built on the coalfields to minimize the cost of transporting coal over long distances. The increasingly surplus rural population flocked to the new manufacturing areas. Canals and other improvements in the transportation infrastructure were made in these regions, which made them attractive to other industries that were not necessarily dependent on coal and thus prompted development in adjacent regions.

Industrialization outside of England began in the mid-19th century in Belgium and northeastern France and spread to Germany, The Netherlands, southern Scandinavia, and other areas in conjunction with the construction of railways. By the 1870s the governments of the European nations had recognized the vital importance of factory production and had taken steps to encourage local development through subsidies and tariff protection against foreign competition. Large areas, however, remained virtually untouched by modern industrial development, including most of the Iberian Peninsula, southern Italy, and a broad belt of eastern Europe extending from the Balkans on the south to Finland and northern Scandinavia.

During the 20th century Europe has experienced periods of considerable economic growth and prosperity, and industrial development has proliferated much more widely throughout the continent; but continued economic development in Europe has been handicapped to a large degree by its multinational character—which has spawned economic rivalries among states and two devastating world wars—as well as by the exhaustion of many of its resources and by increased economic competition from overseas. Many industrial concerns have been deprived of the efficiencies of large-scale production serving a mass market (such as is found in the United States) by governmental protectionism, since this has tended to restrict the potential market for a product to a single country. In addition, enterprise efficiency has suffered from government support and from a lack of competition within a national market area. Within individual countries there have been growing tensions between regions that have prospered and those that have not. This "core–periphery" problem has been particularly acute in situations where the contrasting regions are inhabited by different ethnic groups.

RESOURCES

Mineral resources. With rocks and structures of virtually all geologic periods, Europe possesses a wide variety of useful minerals. Some, exploited since the Bronze Age, are depleted; others have been greatly consumed since the Industrial Revolution. Useful minerals include those that provide energy, ferrous and nonferrous metals and ferroalloys, and those that furnish materials to the chemical and building industries. Europe has a long and commendable prospecting tradition, but, as in the case of North Sea gas and oil, some surprises are still encountered. In relation to the ever-mounting requirements of its economy, however, Europe—the Soviet Union apart—is heavily dependent on mineral imports.

Coal. Europe commands abundant resources of hard and soft coal, which remains of considerable, if declining, importance as a source of power, in smelting minerals, and for its many by-products. Only exceptionally does northern Europe have coal measures of commercial scale, but coal seams are preserved in Hercynian basins throughout the continent, lying diagonally across Britain, Belgium, The Netherlands, France (especially Lorraine), North Rhine-Westphalia and Saarland (West Germany), Upper Silesia (mainly in Poland but also in Czechoslovakia), to the Donets Basin and Urals. There are numerous fields, small but often—as at Komló (near Pécs in Hungary) and in the Arctic fields of Soviet Vorkuta and Norwegian Svalbard—of great locational value. Some, as in southwestern Scotland and southern Belgium, have been worked out or have become uneconomic. Deeper workable seams are sought, in the Ruhr (West Germany), for example, and undersea off Yorkshire (England). Major reserves, encompassing mostly hard deposits of coking, anthracite, and steam coal, lie in the Ruhr, the Pennine fields of England, Upper Silesia, and the Donets Basin. Softer brown coal, or lignite, occurs in the two Germanies, the Chomutov fields of Bohemia, and the Moscow–Tula field.

Petroleum and natural gas. Known petroleum and natural gas reserves are, except in the Soviet Union, wholly disproportionate to Europe's requirements. The Volga-Ural field is the largest in the Soviet Union; Romania's reserves from the Carpathian and sub-Carpathian zones, once the largest in Europe, no longer meet its needs. Many western European countries have located and exploited reserves of petroleum, particularly Norway and the United Kingdom, which have tapped gas and oil from beneath the North Sea bed. In the late 1980s Romania became a leader in extracting oil from the Black Sea.

Uranium. Sources of uranium for use in nuclear reactors have been discovered in many European countries, including France (centred on the Massif Central), Spain, Hungary (the Mecsek Mountains), the Soviet Union (in the Ukraine and Estonia), and, in lesser amounts, in parts of central and eastern Europe.

Iron ores. The largest known iron reserves are the Soviet deposits at Krivoy Rog in the Ukraine, Magnitogorsk, and near Kursk. High-quality ores (of 60 percent iron) from the first two sources have become expensive to mine, but Soviet reserves are more than sufficient for the country and its European associates: the Kursk Magnetic Anomaly, located in the south of the country, has iron-rich quartzites. Deposits in other European countries are small and, except in France and Sweden, inadequate for large-scale heavy industry.

The Kursk Magnetic Anomaly

Ferroalloy metals. The richest ferroalloy deposits occur in the Soviet Union in the emerged shield rocks of the Kola Peninsula (titanium and molybdenum), the Urals, and the Ukraine. Nickel is also mined at Pechenga, in Kola, and at several Ural sites. The southern Urals have also deposits of manganese, required for basic steel manufacture, but these are dwarfed by that at Nikopol, near the Krivoy

Rog iron field, which is the largest and best-located in the world. Other countries have virtually no significant nickel or tin reserves and only small manganese resources. There are chromium deposits of some scale in the Soviet Orsk–Khalilovo region and in Macedonian Yugoslavia, which also contains antimony and molybdenum. Wolframite (for tungsten) is mined from Iberian Hercynian rocks. Norway has molybdenum and titanium workings, and Finland has deposits of titanium, vanadium, and cobalt—valuable and scarce alloys for special steels; the Soviet Union is also an important producer of vanadium.

Nonferrous base metals. With notable exceptions, known European reserves are small, partly as a result of the depletion, for example, of Cornish tin and Swedish copper. Deposits yielding copper, often from copper pyrites, are found in Scandinavia, the southern Urals, and Mediterranean lands. At Bor, Yugoslavia has the largest reserve of copper (low-grade) in Europe; its reserves in lead and, especially, zinc are also high. Mercury is obtained near Krivoy Rog, in Yugoslavia, and in southern Spain. Europe has much bauxite, the principal ore of aluminum, with Greece, Yugoslavia, and Hungary having the largest reserves. Nepheline, an alternative raw material for aluminum, is worked in Kola.

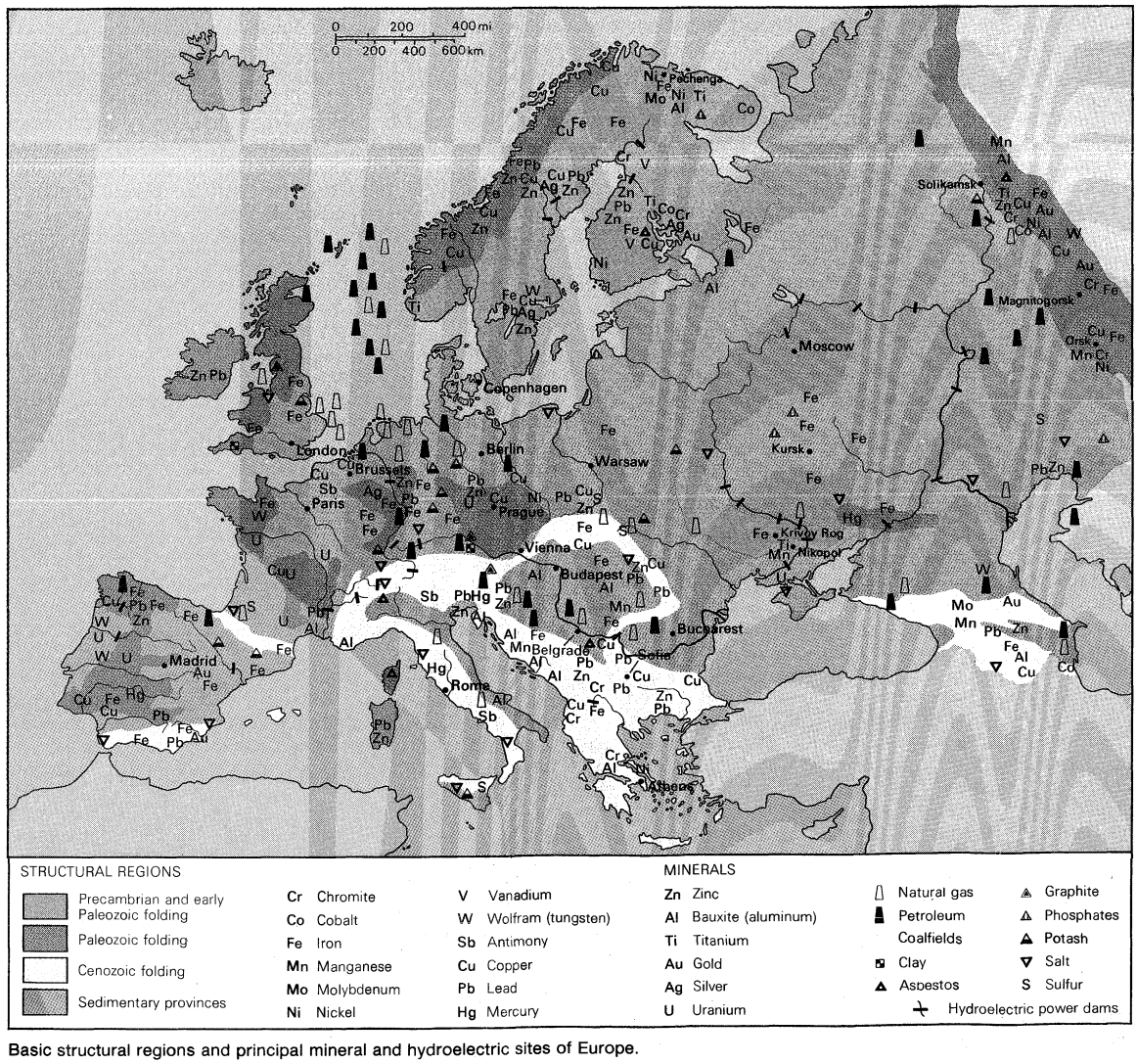
Bauxite reserves

Precious metals. Europe's once widely available reserves of gold appear largely exhausted. Some gold and silver are still produced, mainly in Yugoslavia, Spain, and Sweden.

Nonmetallic deposits. Minerals within this large category are widely available. Clay minerals include fuller's earth—used to cleanse wool and to finish cloth—derived from the decomposition of feldspar; kaolinite, of similar origin, and valuable as china clay, occurs in a pure form in southwestern England: it surpasses coal as a British export; rock salt, important in the chemical industry, occurs widely, much of it being precipitated in such geologically ancient salt lakes as the Soviet Lake Baskunchak (lower Volga), which contains strata 130 feet thick. Europe also has substantial sulfur deposits, and the mining of sulfur in Miocene beds in Sicily gave Italy a virtual monopoly before the opening up of New World deposits in Texas. The carbonate rock, dolomite, like talc, is used as a refractory material, as in lining metal furnaces, and is widespread. Graphite, a crystalline form of carbon used as a lubricant and the basis (with clay) of the "lead" in pencils, is worked in Austria, Czechoslovakia, and England. Nitrates, for fertilizers and explosives, are made from the air electrolytically in England, Norway, and the Soviet Union, and deposits generating potash and phosphate fertilizers are relatively abundant. The Soviet apatite (calcium phosphate) deposits of the Kola Peninsula are the world's largest, as are its potash deposits at Solikamsk, in the Urals. Corundum, a hard abrasive, occurs widely. Building materials for cement and bricks, as well as stone, are abundant, although only regionally available, depending on geologic structure. Particular building stones—marble from central Italy, granite from Norway and Scotland—have localized sources. Except in the Urals, precious stones are rare: these mountains also contain the chief European deposit of asbestos.

Building materials

Water resources. The mountainous and upland areas of Europe collect great amounts of surface water, which supply the rivers and lakes; the lowlands, with lower rainfall, thus receive much of their water from the higher portions of their river basins. In the Mediterranean lands, surface



Basic structural regions and principal mineral and hydroelectric sites of Europe.

water is minimal in summer, exceptions being northern and northwestern Iberia, which gets ample rain; the North Italian Plain, which has Alpine rivers, lakes, springs, and summer rain; and the Apennine zone of Italy. In the European Soviet Union, surface water is relatively abundant in the centre and north but decreases south and southeastward; in the drier regions, however, rivers, collecting water from extensive basins, bring in supplies, and dams on the Volga and Dnepr have created enormous reservoirs.

The increasing water requirements of thermal power stations and industry and, to a lesser extent, domestic needs make the little-populated and industrialized European highlands, which offer surplus water, indispensable to the lowlands. The pollution of water by effluents containing nonoxidizable detergents from urban areas and by those from oil refineries and chemical and metallurgical plants has reached such proportions in, for example, the section of the Rhine below Basel, the Ruhr region, and Lakes Geneva and Garda as to present serious problems and to incur high reclamation costs. In reaction to water shortages, water is, as in the Thames, recycled many times, a practice that leads to the improvement of river water quality.

Europe is relatively well supplied with water, for the water table is normally not far below the surface in the lowlands, and wells and springs are widely available there; underground water supplies (groundwater) that are held particularly in porous rocks are sporadically utilized through the process of pumping. A trend that appears to be growing is to artificially add to supplies of groundwater and thus integrate surface and underground water; nearly half of Sweden's urban water requirements are thus supplied. High capital costs, rather than an actual lack of water, leave some areas of the continent—notably the southeastern European Soviet Union and parts of interior Spain and Turkey—in an arid state. The needs of the major European cities and of the industrial regions involve continuing efforts to collect enough water by impounding surface water, by pumping groundwater, and by encouraging the economy, reuse, and reclamation of water.

Recycling of water

Biological resources. Some reference to plant, animal, and human resources is needed to complement any discussion of European resources. Reference has already been made above to what remains of Europe's plant and animal heritage, supplemented as this is by such vigorous developments as the breeding of livestock to specific purposes and the acclimatization of trees and plants of economic value, which have taken place throughout its history.

The human resources of Europe, since they result from the efforts applied at an ever-rising technical level, are in some respects inexhaustible. Although the cultivation of soil and mining and quarrying for metallic minerals were initiated in prehistoric times, the winning of some resources began only in relatively recent times, in response to new needs and technology. The clearing of woodlands for the plow has continued since the early Middle Ages; the cultivation of the steppe lowlands of the Ukraine and the lower Danube basin commenced only in the late 18th century. Effective drainage, in which the Dutch have excelled, especially during and following the 17th century, has made use of former marshlands. The large-scale mining of coal and iron ore date from the Industrial Revolution. Some industries—many of them concerned with the products of the chemical industry and the refining of aluminum—belong to the 20th century, during which electricity was developed as a form of energy and the internal-combustion engine was developed for use on land, sea, and in the air.

Economic growth stages

The concept of stage, too, helps an understanding of Europe's economic development, for the application to industry and agriculture of modern technology and scientific research has reached different parts of the continent successively. Great Britain, as the home of the Industrial Revolution, stimulated economic change in western, central, and northern Europe. The Soviet Union and the countries of eastern Europe were mostly late starters, and the pace and scale of their industrialization quickened markedly after 1945. The countries of southern Europe, notably northern Italy, also advanced economically following World War II. Europe is thus a highly developed part of the world, although economic development is uneven regionally.

AGRICULTURE

Distribution. Arable land in Europe covers almost 30 percent of the total area, a favourable comparison with both the United States (20 percent) and the Soviet Union (10 percent). Figures for individual countries vary sharply, from about three-fifths of the land in Denmark to less than 3 percent in Norway. Europe's industrialization and urbanization tend to conceal the fact that it is a great producer of cereals, roots, edible oils, fibres, fruit, and livestock and livestock products, accounting for more than 90 percent of the world's rye output, two-thirds of the potato and oats output, and two-fifths of the wheat total.

Europe's climatic range has helped to delineate production areas; thus the vine is commercially grown south of about latitude 50° N, and the olive is restricted to Mediterranean climatic regions. Corn, grown mainly for silage, is an important crop in the lower Danubian lowlands and southern Soviet Union; it appears also in France and Italy. Rice (in northern Italy) and citrus fruits (in Spain, Sicily, and Cyprus) depend on irrigation. The northern countries grow few cereals (mainly oats) and concentrate on animal husbandry, especially cattle and dairying. Grain cultivation is found in the lowland belt that stretches from eastern Great Britain to the Urals. Wheat, in rotation, seeks the better soils, oats and rye the poorer soils and moister lands. Mixed farming—*i.e.*, crop and animal farming—and the use of well-tried crop rotations have become the best economic practice. Viticulture, although widely distributed, is most important in Italy, France, and West Germany, where it is as remunerative as grain cultivation. As to industrial crops, the Soviet Union is the largest producer of flax and hemp, sugar beets (which are also grown widely elsewhere as a rotation crop in the better soils), and sunflower seeds (for edible oil). It is the fifth largest producer of tobacco, which figures importantly also in Bulgaria, Italy, and Macedonian Greece.

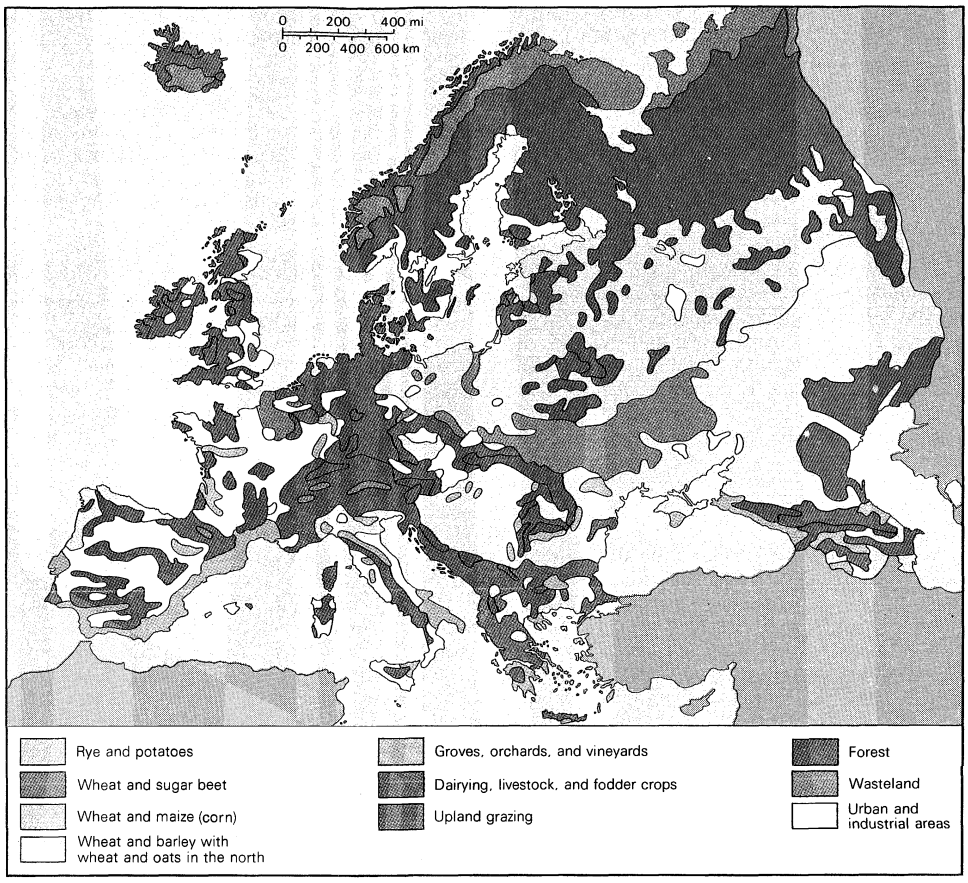
Agricultural organization. There are sharp differences in the way that European agriculture is organized and in its regional efficiency. Collective and state farming is the pattern in the Soviet Union and in most countries of the socialist bloc, but cooperative systems, with or without individual landownership, prevail elsewhere on the continent, with the consolidation of smaller holdings progressing steadily in western Europe. The capital-intensive agriculture of The Netherlands and Great Britain, where yields per acre and per person are high, differs from the extensive Soviet system, where, despite the benefits—notably mechanization—brought by collectivization, labour costs are high and yields low. Less than 3 percent of the working population of the United Kingdom is engaged in agriculture, but one-fourth are so engaged in Yugoslavia. The higher figures indicate high densities of rural population, a shortage of investment capital, and persistent underemployment.

The relative use of fertilizers—very high in The Netherlands and relatively low in Spain and Portugal—hints also at the range of crop productivity.

Irrigated areas, lying mainly in southeastern Spain, the North Italian Plain, and Mediterranean France, are small but disproportionately productive. Long-term prospects for using the irrigation capacity of the lower Danube and Volga are good.

Livestock farming and dairying associated with pigs and poultry is characteristic of European farms, except in the Mediterranean lands, which are better adapted to sheep and goats. Europe produces more than a third of the world's meat, chiefly beef, pork, and bacon, but this is insufficient to meet rising living standards. Domestic production of wool, hides, and leather is also insufficient, except in the Soviet Union, which derives much wool and cotton from its Asian territories. Special features of western European farming include market gardens and the greenhouse production of tomatoes, cucumbers, green vegetables, and flowers for the urban markets. Still another feature is the production of *primeurs:* table fruit, new

Specialty crops of western Europe



Agricultural regions of Europe.

potatoes, vegetables, salad crops, and flowers, produced when prices are high, due to the early spring that visits coastal Brittany, Cornwall, and southern France.

Despite great advances in agronomic science, the hazards of harvest shortfalls due to climate have not been eliminated, and intermittent emergency, as well as regular, claims have to be made on grain-surplus areas overseas. During the late 20th century, harvest shortfalls and increased feed requirements impelled the Soviet Union to import large amounts of grain, especially from the United States and Canada.

INDUSTRY

Mining. Mining provides employment in all countries, although for smaller numbers as mechanization is applied. High-grade iron ores are mined in the Soviet Union at Krivoy Rog and near Kursk and Magnitogorsk and in Arctic Sweden and Norway; these are supplemented by the low-grade minette ores of Lorraine (France) and Luxembourg, low-grade (quarried) Jurassic ores of England, and low-grade Spanish ores. Europe, including the Soviet Union, accounts for about one-third of the world's coal production and nearly four-fifths of its lignite. There was very little increase in coal production during the late 20th century, because European countries have made greater use of other forms of energy, especially oil, nuclear power, natural gas, and hydroelectricity. The chief coal producers are Poland (Upper Silesia), Great Britain, West Germany (the Ruhr and Saarland), and the Soviet Union, where the Donets Basin yields a third of the national output. East Germany is the world's chief source of lignite, mined also in West Germany, Czechoslovakia, and the central European Soviet Union. Many mineral deposits are of only local interest but, as a whole, Europe produces a fair proportion of the world's bauxite, copper, lead, and zinc. Minerals of more than domestic importance are natural gas in The Netherlands; bauxite in Greece, Yugoslavia, and Hungary; petroleum, apatites, and manganese from the Volga-Ural region, Kola, and the Ukraine respectively; and china clay from England.

Heavy industry and engineering. The change from charcoal to coke as fuel in blast furnaces led to the localization of Europe's iron and steel industries on its coalfields to economize transport costs, although imported iron ore, cheap American coal, electric furnaces, and technological efficiency have loosened this tie. Thus, Northumberland and Durham in England, North Rhine-Westphalia, Upper Silesia, and the Donets Basin have their coalfield furnaces and mills, while others are grouped near sources of the ore, as at Krivoy Rog and in Lorraine, or such convenient estuary or port sites as Port Talbot (southern Wales), Genoa, and Dunkirk (France). Europe produces one-half of the world's steel, with the countries of the European Community (EC) and the Soviet Union accounting for about one-third and two-fifths, respectively, of the European output. Europe also produces almost one-third of the world's iron ore. Steel-using industries that make heavy machine tools and mining, smelting, construction, and electrical equipment favour coalfield locations, while shipbuilding and motor-vehicle and aircraft construction show a wider distribution, including new sites.

Coalfield-based industry

Chemical industries. Covering many products, chemical industries have expanded greatly since 1945, partly in relation to hydroelectricity generation and partly as a result of the market-oriented use of refinery by-products. Many heavy chemicals are produced on the coalfields, notably in the Ruhr, where by-products of coke ovens and metallurgical plants are available. Other chemical industries make use of Europe's deposits of salt, potash, phosphates, and sulfur; and the industry has been revolutionized by the increasing production of synthetic rubber, plastics, synthetic fibres, detergents, insecticides, and fertilizers, particularly from petrochemicals.

Manufacturing, lumbering, and fisheries. A wide range of light consumer industries is found throughout Europe, but some countries have acquired reputations for specialty goods, as in the case of English and Dutch bicycles, Swedish and Finnish glass, Parisian perfumes and fashion goods, and Swiss precision instruments and chocolate. The United Kingdom's once-leading textile industry now con-

Specialty goods

centrates on high-quality goods, including many synthetic fibres, of which, together with West Germany, France, and Italy, it is a large producer.

The timber and fisheries extractive industries, now mechanized, are of considerable scale. The Soviet Union, Sweden, and Finland are major producers of softwood and hardwood and exporters of timber, wood pulp, and newsprint. Fishing is a large industry for Norway, Iceland, and the Soviet Union; catches yield not only human food but materials for many subsidiary industries. Fishing is also important in the United Kingdom and France.

Handicrafts and other industries. Of small importance in a continent where the economies of mass production involve standardized production, handicrafts nevertheless survive to serve a wide market, including that of tourists who seek specialty goods. Knitwear and Harris tweed are produced by crofters in the Scottish islands; traditional costumes are made and displayed in many eastern European countries; and custom-made tailoring for men, like dressmaking for women, survives as a supplement to ready-made clothing. Artistic pottery making is another active craft.

Some other European industries fall but uneasily into the preceding categories. Printing and publishing, especially in English, French, German, and Russian, are substantial industries that have worldwide effects, notably in the educational field. Europe is a large producer of pharmaceutical drugs and produces such world-famous beverages as the wines of the west and south, the northern beers, and, not least, whiskey, the status drink from Scotland. Technological and scientific researches are advanced, particularly at such facilities as the European Organization for Nuclear Research (CERN) near Geneva. The outstanding growth industry of tourism—supplementing business, professional, and student travel—brings employment and foreign exchange to many Europeans, especially in the Mediterranean countries, with their combination of sunshine, beaches, scenery, and historical monuments. Europe, with nearly 60 percent of international tourism receipts, is the tourist Mecca of the world.

Growth of tourism

POWER

The message of the Industrial Revolution was that mechanical energy, when it is harnessed to machines, could so supplement human muscle and animal power as to produce revolutionary changes in the scale and pace of factory production. Contemporary Europe, covering less than one-tenth of the inhabited earth, and with only one-seventh its population, uses about one-fourth of the world's energy.

Coal and hydroelectric power. Coal, used to drive steam engines and, as coke, in the smelting of metals, long held the predominant position. During the late 20th century, coal continued to provide energy to coalfield-based industries and was still important for the production of electricity.

Hydroelectricity has been markedly developed where precipitation and landforms provide good opportunities to dam rivers, as in northern and Alpine Europe and the Soviet Union. Norway, for example, derives almost all its electric power from this source; Spain, Portugal, Switzerland, Austria, Sweden, and Yugoslavia derive a large fraction. France has developed power-consuming industries, such as aluminum refining, close to the Alpine and Pyrenean generating sites.

Other power sources. In other countries, hydroelectricity contributes very little, and petroleum and natural gas claim a large share of the energy consumed. By the late 20th century petroleum and natural gas together accounted for one-seventh of the world energy consumption. Natural gas has replaced coal gas in the Soviet Union and Romania and supplements it in Great Britain. Fuel oil is widely used by diesel locomotives and electricity-generating stations and for space heating. Geothermal heat—using underground waters heated by volcanic action—is available in Italy and Iceland, while Ireland, which also lacks both coal and oil, makes efficient use of abundant peat resources. Nuclear-reactor electricity generation, promoted by the European Atomic Energy Community (Euratom) in the EC countries, provides, as in the Soviet Union and eastern European countries, a significant source of electrical energy.

Nuclear-reactor electricity generation

TRADE

Internal and external trade, both by land and by sea, has always been a vigorous part of Europe's economy, and no less so in the late 20th century when Europe faced such strong competitors as the United States and Japan. Trade is made necessary by the regional specialization of production, largely initiated by capitalist enterprise in the past and now predominantly guided by national or even supranational policy decisions. Economic geology partially explains, for example, the localization of Swedish iron and Soviet petroleum production, while climate localizes the production of olive oil and citrus fruits. Europe acquired a central position in modern times in the well-settled Northern Hemisphere, which oceanic and air transport systems still exploit. Simultaneously providing large managerial, market, and labour-force attractions, Europe inevitably attracts extra-European traders, with its ever more sophisticated industry producing outstanding exports and its large importation of petroleum products, metals, raw materials, and foodstuffs.

Within the continent, there has been a distinction between the general trade policy of western Europe and that of the socialist bloc. Prior to the late 1960s the Soviet Union and the eastern European countries adhered to the doctrine of economic self-sufficiency with more interregional than international trade. In the late 1960s and the '70s these trading patterns began to change. Improved relations between the East and the West enabled the socialist countries to meet an increased amount of their technological and agricultural needs with imports from Western countries. The nations of western Europe, on the other hand, have always relied heavily on international trade.

Europe plays the leading role in world commerce, making up more than one-half the total of world exports and imports. The bulk of this trade is carried on by the Western countries, which own almost one-fourth of the world's oceangoing tonnage. Most of the European countries long held political dependencies overseas where they created captive markets, and this imperialist trading momentum persists. Common Market countries have former colonial territories as associate members, and, similarly, the Commonwealth nations engage in much trade, now strictly competitive, with the United Kingdom, although the latter's accession to the Common Market in 1973 resulted in a decline in the proportion of British trade with the Commonwealth.

The international financial crisis

One of the continuing international difficulties that Europe has faced concerns currency and fluctuating exchange rates, as the currencies of Britain, France, and the United States have come under strain, while those of West Germany, The Netherlands, and Switzerland have grown harder, or relatively more valuable. European financiers play an important world role, as do a variety of such finance-related industries as banking, insurance, and shipping.

Internal trade. Within each European country a wide variety of goods is moved continually from ports and production centres to urban markets. Miscellaneous home-produced goods are also traded to consumer centres. Imported goods include fuels, tropical foodstuffs and drinks, raw materials, textile fibres, metals and metallic ores, and a wide range of manufactured goods.

Trading blocs. Active trading within groups of countries that have associated primarily for this purpose and to rationalize and so increase the profitability of their national economies has advanced. The policies of the EC and the Council for Mutual Economic Assistance (CMEA, or Comecon) have been directed toward economic specialization in increasingly interdependent member countries. West Germany supplies coking coal and chemicals to France, which provides Belgium with iron ore from Lorraine. Steel is moved to extranational markets, and Dutch natural gas is piped to France, Belgium, and West Germany. Specialty foodstuffs—wines, cheeses, spring vegetables, and fruit—find an enlarged market far beyond

their production centres, as do such manufactured items as fashion goods, automobiles, and major household appliances. Comecon countries direct their trade largely to each other, with the Soviet Union preeminent as a producer and a major market. The Soviet Union supplies petroleum, manganese, iron, and chrome ores, as well as cotton and other textile fibres; it receives machinery, textiles, and consumer goods.

The European Free Trade Association (EFTA) has also encouraged trade between its members, who exchange such complementary, rather than similar, products as Swedish and Finnish timber and Swiss watches and food products. In 1977 a free-trade agreement went into effect between the Common Market and the EFTA. The agreement eliminated tariffs on most industrial goods originating in the member countries, thereby increasing trade between the countries in the two blocs.

Trade between the West and the socialist bloc increased markedly during the late 20th century. Soviet natural gas was sold to Italy, France, and West Germany, and Western markets were also used for the sale of gold and diamonds in exchange for ships, machinery, and chemicals. European socialist countries supplied Soviet automobiles, canned salmon and caviar, vodka, Polish bacon, Czech glass, and Hungarian and Yugoslav wines.

Intra-Soviet trade. Given its continental scale, the regional trade of the Soviet Union, carried largely by a heavily worked railway system, supplemented by pipelines and an elaborate waterway system, deserves special notice. From the south, grain, meat, vegetable oils, sugar, tobacco, wine, and fruits are moved to the central and northern Soviet Union, where the consumption of such items exceeds local production; dried fish and salt are carried up the Volga. Timber is moved from northern, including Ural, areas to largely unforested southern regions, and Donets Basin coal is shipped by canal to Volgograd.

External trade. A major part of the external trade of European countries is with each other since, with regional specialization, dense populations, and relatively high standards of living, they provide strong markets. For the Common Market countries, as well as for those of northern Europe (EFTA), this trade proportion is very high. In the socialist bloc about three-fourths of the foreign trade is intra-Comecon. There is, nevertheless, a substantial amount of trading among EC, EFTA, and Comecon members, and, especially in the United Kingdom and West Germany, a vigorous two-way trade with the United States is conducted. European trade also extends to all other parts of the world, including the developing countries, where—in exchange for manufactured products—vital supplies of energy, raw materials, metals, ores, and foodstuffs are obtained.

North American trade links

The extracontinental exports of Europe include machine tools, automobiles, aircraft, chemicals (including pharmaceutical drugs), and such consumer items as clothing, textiles, books, expert services, and works of art. Western Europe depends heavily on imported petroleum from the Middle East, Algeria, and Libya and on many raw materials and metals. Europe imports much natural rubber, tea, coffee, cacao, cane sugar, vegetable oilseeds, tobacco, and fruit—fresh, canned, and dried—although it has attempted to lessen its dependence on imported agricultural products with greater domestic production and the manufacture of synthetic substitutes for vegetable fibres.

TRANSPORTATION

Roads. The European Soviet Union is, compared with western Europe—where motorways provide fast movement for commerce and travel—relatively deficient in engineered motor roads. Motorable roads have become more widely available; those of Spain and Ireland have particularly improved, and road tunnels now supplement railway tunnels beneath the Alpine passes. Animal transport has minimal importance yet survives locally: the horse-drawn cart may still be seen in east central Europe; and the ox-drawn plow and the loaded ass, mule, and donkey—surefooted in rough, hilly country—are still used in parts of southern Europe. In such regions with long, snowy winters as the Soviet North, the dog- or reindeer-drawn sled is used.

The role of animals in transport

Railways. Railways link European ports with their hinterlands and fan out from capitals and major cities to points on the international frontiers where they meet the railway system of their neighbours. In some cases—notably from France to Spain, and the Soviet Union to Poland—this involves a change of gauge. Underground and suburban railways also play an indispensable role for metropolitan commuters. Railways permit passage between the western and eastern European extremities but not quite to the extreme north; they have lost some of their passengers and freight to the automobile, coach, and truck, and many uneconomic local lines have been closed. Even so, rail services have notably improved with the use of electrified track or diesel locomotives, faster intercity passenger trains, and container freight trains. Railways remain all-important in the Soviet Union.

Waterways and pipelines. Seaports have been modernized and enlarged to deal more efficiently with the increasing size of ships and volume of oceanic trade. Even landlocked Switzerland has seagoing ships that use Dutch port facilities. The United Kingdom, Norway, and Greece also hold large freighter tonnages for hire.

Inland waterway transport, slow but cheap, is regionally important for the carriage of heavy and bulky commodities. The best waterways—the Rhine below Rheinfelden and the Danube below Belgrade—can carry 1,500-ton barges. The navigable Rhine has the legal status of an international waterway open to all users. Other rivers and canals are usable by smaller vessels. The Volga, however, is a valuable waterway linking Moscow with Caspian ports and, via the Volga–Don Shipping Canal, gives water access to the Donets Basin.

Barge carriers

Giant tankers, up to and beyond 300,000 gross registered tonnage and too deep in draught for most seaports, deliver their cargoes by pipelines that—for petroleum, natural gas, and water—provide the cheapest overland form of transport. They have been built in the United Kingdom for North Sea gas and oil; in France, Spain, and Italy for North African oil; and within and beyond the Soviet Union, whose Friendship Pipeline carries crude oil from the Volga-Ural field to eastern European refineries.

Airways. Air services between principal European cities and to all parts of the world are extensively organized. Airports at London, Frankfurt am Main, Paris, Stockholm, Amsterdam, and Moscow stand out as those of first importance. Passengers, mail, and commodities of high value in relation to their weight—such as gold and early spring flowers—make use of air transport.

(W.G.E./T.M.P.)

EUROPEAN GEOGRAPHIC FEATURES OF SPECIAL INTEREST

Landforms

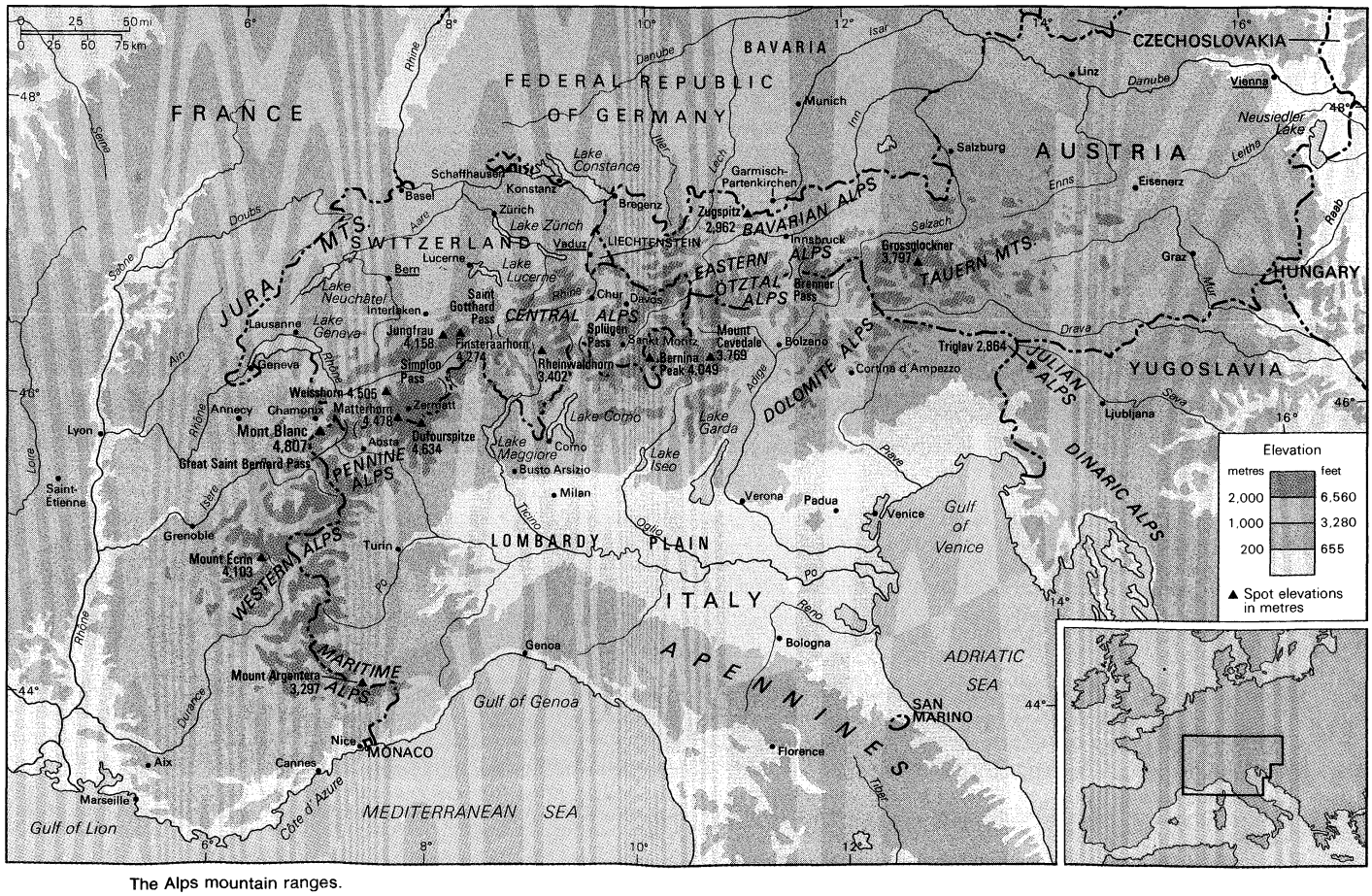
THE ALPS

The Alps are a small segment of a discontinuous mountain chain that stretches from the Atlas Mountains of North Africa across southern Europe and Asia to beyond the Himalayas. They extend north from the subtropical Mediterranean coast near Nice, Fr., to Lake Geneva before trending east-northeast to Vienna (at the Vienna Woods). There they touch the Danube River and meld with the adjacent plain. The Alps form part of six nations: France, Italy, Switzerland, West Germany, Austria, and Yugoslavia; only Switzerland and Austria can be considered true Alpine nations. Some 750 miles (1,200 kilometres) long and more than 125 miles wide at their broadest point between Garmisch-Partenkirchen, W.Ger., and Verona, Italy, the Alps are the most prominent of western Europe's physiographic regions.

Though they are not as high and extensive as other mountain systems uplifted during the Tertiary period—such as the Himalayas and the Andes and Rocky mountains—they are responsible for major geographic phenomena. The Alpine crests isolate one European region from another and are the source of many of Europe's major rivers, such as the Rhône, Rhine, Po, and numerous tributaries of the Danube. Thus, waters from the Alps ultimately reach the North, Mediterranean, Adriatic, and Black seas. Because of their arclike shape, the Alps separate the marine west-coast climates of Europe from the Mediterranean areas of France, Italy, and Yugoslavia. Moreover, they create their own unique climate based on both the local differences in elevation and relief and the location of the mountains in relation to the frontal systems that cross Europe from west to east. Apart from tropical conditions, most of the other climates found on the Earth may be identified somewhere in the Alps, and contrasts are sharp.

A distinctive Alpine pastoral economy that evolved through the centuries has been modified since the 19th century by industry based on indigenous raw materials, such as the industries in the Mur and Mürz valleys of southern Austria that used iron ore from deposits near Eisenerz. Hydroelectric power development at the end of the 19th and beginning of the 20th centuries, often involving many different watersheds, led to the establishment in the lower valleys of electricity-dependent industries, manufacturing such products as aluminum, chemicals, and specialty steels. Tourism, which began in the 19th century in a modest way, has become, since the end of World War II, a mass phenomenon. Thus, the Alps have become a summer and winter playground for millions of European urban dwellers and annually attract tourists from around the world. Because of this enormous human impact on a fragile physical and ecological environment, the Alps are the most threatened mountain system in the world.

Physical features. *Geology.* The Alps emerged during the Alpine orogeny, an event that began about 70 million years ago as the Mesozoic era was drawing to a close. A broad outline helps to clarify the main episodes of a complicated process. At the end of the Paleozoic era, about 245 million years ago, eroded Hercynian mountains, similar to the present Massif Central in France and Bohemian Massif embracing parts of West and East Germany, Austria, Poland, and Czechoslovakia, stood where the Alps are now located. A large landmass, formed of crystalline rocks and known as Tyrrhenia, occupied what is today the western Mediterranean basin, whereas much of the



The Alps mountain ranges.

rest of Europe was inundated by a vast sea. During the Mesozoic (245 to 66.4 million years ago) Tyrrhenia was slowly leveled by the forces of erosion. The eroded materials were carried southward by river action and deposited at the bottom of a vast ocean known as the Tethys Sea, where they were slowly transformed into horizontal layers of rock composed of limestone, clay, shale, and sandstone.

During the middle Tertiary period, about 44 million years ago, relentless and powerful pressures from the south first formed the Pyrenees and then the Alps, as the deep layers of rock that had settled into the Tethys Sea were folded around and against the crystalline bedrock and raised with the bedrock to heights approaching the present-day Himalayas. These tectonic movements lasted until nine million years ago. Tyrrhenia sank at the beginning of the Quaternary period, about 1.6 million years ago, but remnants of its mass, such as the rugged Estéral region west of Cannes, are still found in the western Mediterranean. Throughout the Quaternary period, erosive forces gnawed steadily at the enormous block of newly folded and upthrust mountains, forming the general outlines of the present-day landscape.

Glaciation

The landscape was further modeled during the Quaternary by Alpine glaciation and by expanding ice tongues, some reaching depths of nearly one mile (1.6 kilometres), that filled in the valleys and overflowed onto the plains. Amphitheatre-like cirques, arête ridges, and majestic peaks such as the Matterhorn and Grossglockner were shaped from the mountaintops; the valleys were widened and deepened into general U-shapes, and immense waterfalls, like the Staubbach and Trümmelbach falls in the Lauterbrunnen Valley of the Bernese Alps, poured forth from hanging valleys hundreds of feet above the main valley floors; elongated lakes of great depth such as Lake Annecy in France, Lake Constance bordering Switzerland, West Germany, and Austria, and the lakes of the Salzkammergut in Austria filled in many of the ice-scoured valleys; and enormous quantities of sands and gravels were deposited by the melting glaciers, and landslides—following the melting of much of the ice—filled in sections of the valley floors. The hills east of Sierre in the Rhône valley are an example of this last phenomenon, and they mark the French–German language divide in this area.

When the ice left the main valleys, there was renewed river downcutting, both in the lateral and transverse valleys. The river valleys have been eroded to relatively low elevations that are well below those of the surrounding mountains. Thus, Aosta, Italy, in the Pennine Alps and Sierre, Switz., look up to peaks that tower a mile and a half above them. In the valley of the Arve River near Mont Blanc, the difference in relief is more than 13,100 feet.

Glaciation therefore modified what otherwise would have been a harsher physical environment: the climate was much milder in the valleys than on the surrounding heights, settlement could be established deeper into the mountains, communication was facilitated, and soils were inherently more fertile because of morainic deposits. Vigorous glacial erosion continues in modern times. Many hundreds of square miles of Alpine glaciers, such as those in the Ortles and Adamello ranges and such deep-valley glaciers as the Aletsch Glacier near Brig, Switz., are still found in the Alps. The summer runoff from these ice masses is instrumental in filling the deep reservoirs used to generate hydroelectricity.

Physiography. The Alps present a great variety of elevations and shapes, ranging from the folded sediments forming the low-lying pre-Alps that border the main range everywhere except in northwestern Italy to the crystalline massifs of the inner Alps that include the Belledonne and Mont Blanc in France, the Aare and Gotthard in Switzerland, and the Tauern in Austria. From the Mediterranean to Vienna, the Alps are divided into Western, Central, and Eastern segments, each of which consists of several distinct ranges.

Divisions of the Alps

The Western Alps trend north from the coast through southeastern France and northwestern Italy to Lake Geneva and the Rhône valley in Switzerland. Their forms include the low-lying arid limestones of the Maritime Alps near the Mediterranean, the deep cleft of the Verdon Canyon in France, the crystalline peaks of the Mercantour Massif, and the glacier-covered dome of Mont Blanc, which at 15,771 feet (4,807 metres) is the highest peak in the Alps. Rivers from these ranges flow west into the Rhône and east into the Po.

The Central Alps occupy an area from the Great St. Bernard Pass east of Mont Blanc on the Swiss-Italian border to the region of the Splügen Pass north of Lake Como. Within this territory are such distinctive peaks as the Dufourspitze, Weisshorn, Matterhorn, and Finsteraarhorn, all 14,000 feet high. In addition, the great glacial lakes—Como and Maggiore in the south, part of the drainage system of the Po; and Thun, Brienz, and Lucerne (Vierwaldstättersee) in the north—fall within this zone.

The Eastern Alps, consisting in part of the Rätische range in Switzerland, the Dolomite Alps in Italy, the Bavarian Alps of West Germany and Austria, the Tauern Mountains in Austria, and the Julian Alps in Yugoslavia, are synonymous with a northerly and southeasterly drainage pattern. The Inn, Lech, and Isar rivers in West Germany and the Salzach and Enns in Austria flow into the Danube north of the Alps, while the Mur and Drau (Austria) and Sava (Yugoslavia) rivers discharge into the Danube east and southeast of the Alps. Within the Eastern Alps in Italy, Lake Garda drains into the Po, whereas the fast-flowing Adige, Piave, Tagliamento, and Isonzo pour into the Gulf of Venice.

Differences in relief within the Alps are considerable. The highest mountains, composed of autochthonous crystalline rocks, are found in the west in the Mont Blanc massif and also in the massif centring on Finsteraarhorn (14,022 feet) that divides the cantons of Valais and Bern. Other high chains include the crystalline rocks of the Mount Blanche nappe—which includes the Weisshorn (14,780 feet)—and the nappe of Monte Rosa Massif, sections of which mark the frontier between Switzerland and Italy. Farther to the east, Bernina Peak is the last of the giants over 13,120 feet (4,000 metres). In Austria the highest peak, the Grossglockner, reaches only 12,457 feet; West Germany's highest point, the Zugspitze in the Bavarian Alps, only 9,718 feet; and the highest point of Yugoslavia and the Julian Alps, Triglav, only 9,396 feet. Some of the lowest areas within the Western Alps are found at the delta of the Rhône River where the river enters Lake Geneva, 1,220 feet. In the valleys of the Eastern Alps north of Venice, elevations of only about 300 feet are common.

Climate. The location of the Alps, as well as the great variations in their elevations and exposure, give rise to extreme differences in climate, not only among separate ranges but also within a particular range itself. Because of their central location in Europe, the Alps are affected by four main climatic influences: from the west flows the relatively mild, moist air of the Atlantic; cool or cold polar air descends from northern Europe; continental air masses, cold and dry in winter and hot in summer, dominate in the east; and, to the south, warm Mediterranean air flows northward. Daily weather is influenced by the location and passage of cyclonic storms and the direction of the accompanying winds as they pass over the mountains.

Temperature extremes and annual precipitation are related to the physiography of the Alps. The valley bottoms clearly stand out because generally they are warmer and drier than the surrounding heights. In winter, nearly all precipitation above 5,000 feet is in the form of snow, and depths from 10 to 33 feet or more are common. Snow cover lasts from approximately mid-November to the end of May at the 6,600-foot level, blocking the high mountain passes; nevertheless, relatively snowless winters can occur. Mean January temperatures on the valley floors range from 23° to 39° F (−5° to 4° C) to as high as 46° F (8° C) in the mountains bordering the Mediterranean, whereas mean July temperatures range between 59° and 75° F (15° and 24° C). Temperature inversions are frequent, especially during autumn and winter, and the valleys often fill with fog and stagnant air for days at a time. At those times the levels above 3,300 feet can be warmer and sunnier than the low-lying valley bottoms. Winds can play a prominent role in daily weather and microclimatic conditions.

The foehn

A foehn wind can last from two to three days and blows either south–north or north–south, depending on the tracking of cyclonic storms. The air mass of such a wind is cooled adiabatically as it passes upward to the mountain crests, which precipitates either rain or snow and retards the rate of cooling. When this drier air descends on the lee side, it is adiabatically warmed by compression at a constant rate and therefore has a higher temperature at the same altitude than when it began its upward flow. Snow in the affected areas disappears rapidly.

Avalanches, one of the great destructive forces of nature, are an ever-present danger during the period from late November to early June. Though occurring wherever there are high mountains, open slopes, and heavy snowfalls, avalanches are a greater hazard in the Alps than in other mountain ranges because of the relatively high population density and the expansion of winter tourism. Avalanches not only cause widespread damage but, by carrying down large quantities of rock from the mountain slopes to the valley floors, also are significant agents of erosion. Most avalanches follow well-defined paths, but much of the fear of avalanches is related to the difficulty of predicting where and when they will strike.

Plant and animal life. Several vegetation zones that occur in the Alps reflect differences in elevation and climate. A variety of species of deciduous trees grow on the valley floors and lower slopes; these include linden, oak, beech, poplar, elm, chestnut, mountain ash, birch, and Norway maple. At higher elevations, however, the largest extent of forest is coniferous; spruce, larch, and a variety of pine are the main species. For the most part, spruce dominance reaches its upper limit at approximately 7,200 feet in the Western Alps. Better able to resist conditions of cold, lack of moisture, and high winds, larch can grow as high as 8,200 feet and are found interspersed with spruce at lower elevations. At the upper limits of the forests are hardy species such as the Arolla pine that generally do not grow below the 5,000-foot level; this slow-growing tree can live for 350–400 years and in exceptional cases up to 800 years. Its wood, strongly impregnated with resin, decays very slowly and was formerly prized for use in the construction of chalets. The areas of Arolla pine have been so reduced that cutting the trees is strictly controlled. Above the tree line and below the permanent snow line, a distance of about 3,000 feet, are areas eroded by glaciation that in places are covered with lush Alpine meadows. There sheep and cows are grazed during the short summer, a factor that has helped lower the upper limits of the natural forest. These distinctive mountain pastures—called *alpages,* from which both the names of the mountain system and the vegetational zone are derived—are found above the main and lateral valleys; the spread of invasive weeds, pollution from animal wastes, and erosion from ski-related development limit their carrying capacity. In the southern reaches of the Maritime Alps and the southern Italian Alps, Mediterranean vegetation dominates, with maritime pine, palm, sparse woodland, and agave and prickly pear evident.

The *alpages*

A few species of animals have adapted well to the higher mountains. Bears have vanished, but the ibex, which like the chamois is endowed with extraordinary nimbleness, was saved by Italian royal game preserves. Marmots hibernate in underground galleries. The mountain hare and the ptarmigan—a grouse—assume a white coat for winter. Several national parks amid the ranges ensure preservation of the native fauna.

Human impact on the Alpine environment. The early travelers to the Alps were greatly inspired by the pristine beauty of what they saw, and from their inspiration sprang the modern popularity of the Alpine region. With popularity, however, came growth; and the impact of so many people has caused a steady degradation of the Alpine environment since the mid-20th century. This has resulted in air of poorer quality; water pollution in rivers and lakes; a rise in noise pollution; slope erosion caused by the construction of ski slopes and roads; dumping, often indiscriminately, of solid and organic waste; erosion from the quarrying of rock, sand, and gravel for construction; and forests weakened by acid rain. Slowly, the unique landscape and flora of the Alps that so inspired the early travelers is being irrevocably altered.

Most conspicuous, perhaps, is the obvious transformation of the landscape. The main river valleys have been converted into linear conurbations of concrete and asphalt; and, in order to accommodate the expanding tourist trade, many villages in the higher lateral valleys have taken on the character of lowland suburbs. A highly visible result of this growth is the serious decline in air quality. Pollution from factories adds to that from home heating and motor vehicle exhausts, the situation aggravated by temperature inversions and weather conditions that often produce little wind. Many of the larger Alpine cities experience severe local air pollution, and some of the valleys can be filled with impure air for weeks at a time.

Pollution

The people. *Settlement.* Humans have been living in the Alps since Paleolithic times, 60,000 to 50,000 years ago. They hunted game and left their artifacts in various sites from the Vercors near the Isère valley in France to the Lieglhohle above Taupliz in Austria. After the retreat of the Alpine glaciers, 4,000 to 3,000 years ago, the valleys were inhabited by Neolithic peoples who lived in caves and small settlements, some of which were built on the shores of the Alpine lakes. Sites have been discovered near Lake Annecy, along the shores of Lake Geneva, in the Totes Mountains in Austria, and in the Aosta and Camonica Valleys in Italy. The latter valley is noted for some 20,000 rock engravings that leave an invaluable picture of more than 2,000 years of habitation.

From 800 to 600 BC Celtic tribes attacked the Neolithic encampments and forced their inhabitants into the remote valleys of the Alps. In the west the area around the juncture of France, Switzerland, and Italy was occupied by the Celts; the modern urban centres of the region, including Martigny, Switz., Aosta, Italy, and Grenoble, Fr., owe their origin to these people. The Celts also penetrated the valleys of Graubünden canton in eastern Switzerland, but the great centre of Celtic culture was found at Hallstatt, the site of a small settlement in Upper Austria. Because of rich archaeological finds there the name Hallstatt has become synonymous with the late Bronze and early Iron ages in Europe, a period dating from about 1000 to 500 BC. The Celts began to open the high Alpine passes for trade routes.

The Romans enlarged the old Celtic villages and built many new towns both in the valleys leading up to the Alps and within the Alps themselves. Villa Aniciaca (modern Annecy, Fr.), Octodurus (Martigny), Augusta Praetoria (Aosta), and Virunum (Zollfeld, Austria) flourished under Roman rule. The Romans improved water supplies and constructed arenas and theatres, the best preserved of which is in Aosta. Control of the Alpine passes was the key to Roman expansion, and they were enlarged from trails to narrow roads. The passes that linked the Roman outposts (*e.g.,* Great St. Bernard, Splügen, Brenner, and Plöcken) were particularly important. The first of the "barbaric" incursions took place in AD 259, and by 400 Roman control of the Alps had disintegrated.

The lands of the Romanized Celts were occupied by Germanic tribes that included the Burgundians, Alemanni, and Lombards. During the 8th and 9th centuries the Alpine lands became part of Charlemagne's Holy Roman Empire. The Treaty of Verdun (843) divided the empire among Charlemagne's grandsons, and in 888 further partition resulted in the basic linguistic differences that have endured until the present. The unity that was imposed on the Alps by the Celts, Romans, and barbarians disappeared during the Middle Ages. For the most part, each valley lived apart and isolated from its neighbours. Much of the history of Alpine peoples after the Roman domination, mirroring that of Europe as a whole, was characterized by an expedient and continuous shifting of religious and political alliances. The isolation of the Alpine peoples was broken by the Industrial Revolution and the coming of the railways that penetrated the Alps via great tunnels.

Partitions of Alpine lands

Languages. French is spoken in the Western Alps, including the Swiss cantons of Vaud and Valais, and in the northwestern Italian region of the Valle d'Aosta. Ostensibly bilingual, the Valle d'Aosta has not been able to resist

the impact of Italianization, and the use of French in daily affairs is confined to certain of the lateral valleys. Italian is spoken in the Central and Eastern Alps of Italy and in the Swiss canton of Ticino. The German language is used throughout the Central and Eastern Alps of Switzerland, Germany, and Austria, as well as in the Alto Adige region of Italy (before World War I the Südtirol area of Austria). There are pockets of Ladin and Friulian peoples in the Eastern Alps of northeastern Italy, and Slovenian is spoken in Yugoslavia and the adjacent Alpine border regions with Italy and Austria. Roman Catholicism is the main religion throughout the Alps, although there are regions that are predominantly Protestant, such as the Swiss cantons of Vaud and Bern. The Swiss canton of Graubünden reflects the diversity of languages and religion in the Alps, where some 45 percent of its population is Protestant and 50 percent Catholic; 60 percent speak German, about 15 percent Italian, and 20 percent Romansh. Added to the mixture of indigenous languages is the babel created by the variety of foreign seasonal workers, without whom the tourist industry, especially in Switzerland, would collapse.

The economy. *Agriculture.* Before the mid-19th century the economic basis of the Alps was predominantly agricultural and pastoral. Though since then there has been widespread abandonment of farms, especially in the high valleys of France and Italy and in western Austria, agriculture still survives in favoured locations both in the main and lateral valleys. The hot and dry Rhône valley in Switzerland, between Sierre and Martigny, is an intensive area of irrigated fruit and vegetable cultivation, and both the valley floor and slopes of the mountains have extensive vineyards from which excellent wines are made. Above Visp are some of the highest vineyards in the world, reaching more than 4,250 feet. Other regions of viticulture include the Alto Adige region in northern Italy, Ticino, and the southern regions of the Alps. Villagers in such locations as Chandolin in the Swiss Anniviers Valley—which at 6,561 feet is the highest settlement inhabited year-round in the Alps—cut grass for feeding dairy cows, but most of the agriculture and pastoralism in the high valleys exists as hobby farming or second-income enterprises.

Viticulture

Mining and manufacturing. The mainstay of the modern Alpine economy is a combination of mining and quarrying, manufacturing, industries, and tourism. Mining has been carried out since Neolithic times and is still significant in the Erzberg of Austria, where iron has been extracted from the mountain since the Middle Ages. Near Cluse, in the pre-Alps of Haute-Savoie not far from Geneva, a region of watch making, screw cutting, component manufacturing, and related industries emerged in the first quarter of the 19th century and evolved into one of the most concentrated industrial locations of its type in the world. Large steel mills were located in Aosta and in the Mur and Mürz valleys because of local supplies of iron and coal. In addition, pulp and paper plants that utilized the Alpine forests were established in the Eastern Alps of Austria. With the development of hydroelectricity in the late 19th and 20th centuries, heavy metallurgical and chemical industries were attracted to the major transverse valleys of France, southern Switzerland, and western Austria. Later, factories producing such consumer products as textiles (in the Rhine valley of Austria) and sporting goods (the Annecy area in France) were established. One result of this industrialization was the depopulation of the small villages in the lateral valleys, an occurrence that was partially stemmed by the emergence of the tourism boom after 1960. Many of the early industrial enterprises are no longer viable because of obsolescence, foreign competition, the high cost of transporting raw materials from coastal ports to interior valley locations, or—as is the case with the steel plant in Aosta—because indigenous raw materials have been exhausted. The remaining plants have had to modernize, rationalize, restructure, and develop new products in order to remain competitive in world markets.

Tourism. The most significant economic change for the Alps has been the development of mass tourism since World War II. Tourism in the Alps is a risky business: capital investment can be considerable, whereas the season in which to recoup expenditure is short and can be disrupted by economic difficulties in neighbouring countries or by a lack of snow in winter and cool, rainy weather in summer. Furthermore, there is fierce competition to attract tourists, not only among the different Alpine countries but also among the resorts within each country. There are some 600 ski resorts in the Alps, with more than 270 in Austria alone. Nevertheless, winter and summer tourism have injected enormous sums of money into the economies of the Alpine nations, a development that has been especially beneficial to the remote villages of the high lateral valleys. Employment opportunities in the service sector have increased substantially, taking up the slack caused by a decline in agricultural and industrial employment.

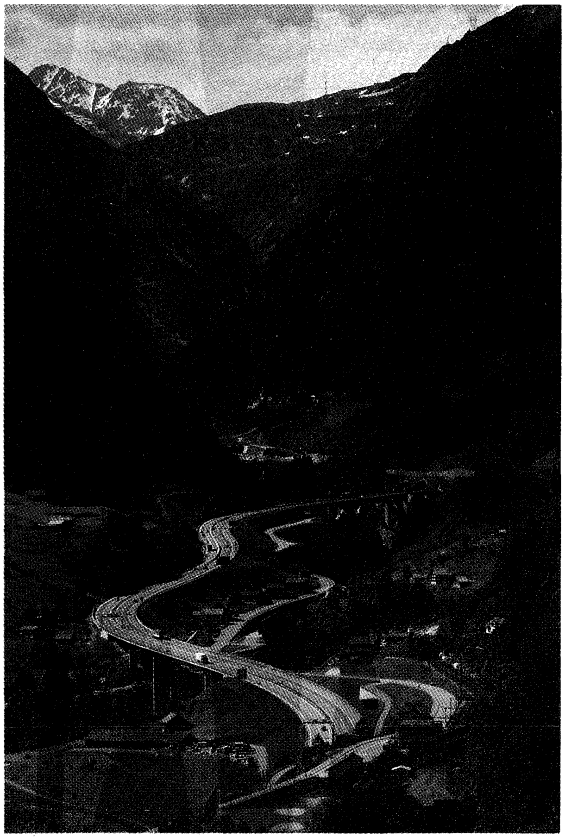
Transportation. The rugged and steep terrain of the Alps long was a barrier to transportation. Beginning in Celtic times, however, and continuing into the present, mountain passes have served as communication links between otherwise isolated valleys; the passes have evolved from simple paths to paved, multilane highways. Such settlements as Chur in eastern Switzerland, a focal point for the numerous passes in the region, have been inhabited for more than 5,000 years. Andermatt, in south central Switzerland, grew in a similar manner.

Role of mountain passes

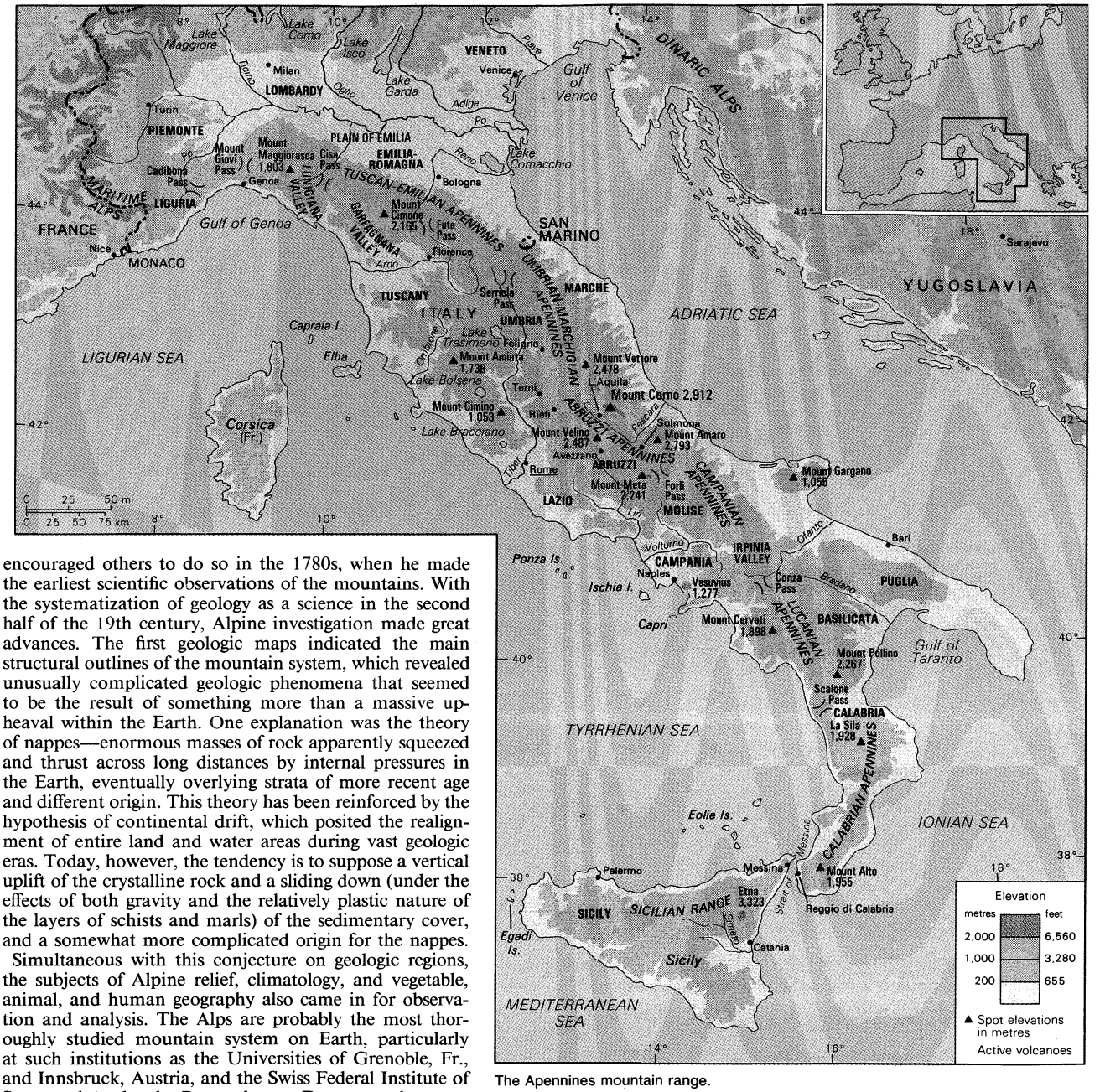
The advent of rail and later road transportation and the accompanying improvements in road-building techniques have ended the isolation of most areas of the Alps. Tunnels—and road tunnels in particular—which allow huge numbers of people to pass under the great Alpine massifs at all times of the year, have had the greatest impact: by facilitating such a steady onslaught of motor vehicles and people, they not only have made possible the tremendous growth in tourism in the 20th century but also have become a major contributing factor in the degradation of the Alpine environment.

Study and exploration. Records of ascents of various peaks in the Alps date from at least as early as the 14th century, and, in the late 18th and the 19th centuries, the interest in this activity created a vogue for serious mountaineering that began in the Alps and spread throughout the world. Horace Bénédict de Saussure, a professor at the University of Geneva, made ascents of the peaks and

Aubrey Diem



Highway to St. Gotthard tunnel and pass, Wassen, Switz.



The Apennines mountain range.

encouraged others to do so in the 1780s, when he made the earliest scientific observations of the mountains. With the systematization of geology as a science in the second half of the 19th century, Alpine investigation made great advances. The first geologic maps indicated the main structural outlines of the mountain system, which revealed unusually complicated geologic phenomena that seemed to be the result of something more than a massive upheaval within the Earth. One explanation was the theory of nappes—enormous masses of rock apparently squeezed and thrust across long distances by internal pressures in the Earth, eventually overlying strata of more recent age and different origin. This theory has been reinforced by the hypothesis of continental drift, which posited the realignment of entire land and water areas during vast geologic eras. Today, however, the tendency is to suppose a vertical uplift of the crystalline rock and a sliding down (under the effects of both gravity and the relatively plastic nature of the layers of schists and marls) of the sedimentary cover, and a somewhat more complicated origin for the nappes.

Simultaneous with this conjecture on geologic regions, the subjects of Alpine relief, climatology, and vegetable, animal, and human geography also came in for observation and analysis. The Alps are probably the most thoroughly studied mountain system on Earth, particularly at such institutions as the Universities of Grenoble, Fr., and Innsbruck, Austria, and the Swiss Federal Institute of Snow and Avalanche Research near Davos; yet they continue to present hosts of complex and evolving scientific problems. (P.V./A.Di.)

APENNINES

The Apennines (Italian: Appennino), a series of mountain ranges bordered by narrow coastlands, form the physical backbone of peninsular Italy and have had considerable influence on the human geography of that nation. From Cadibona Pass in the northwest, close to the Maritime Alps, they form a great arc, which extends as far as the Egadi Islands to the west of Sicily. Their total length is approximately 870 miles (1,400 kilometres), and their width ranges from 25 to 125 miles. Mount Corno, 9,554 feet (2,912 metres), is the highest point of the Apennines proper on the peninsula. The range follows a northwest–southeast orientation as far as Calabria, at the southern tip of Italy; the regional trend then changes direction, first toward the south and finally westward.

The Apennines are among the younger ranges of the Alpine system and, geologically speaking, are related to the coastal range of the Atlas Mountains of North Africa. Similarities have also been observed with the Dinaric Alps, which extend through Yugoslavia and Greece. The nearby islands of Sardinia and Corsica, on the other hand, are dissimilar to the Apennines, their granitic rock masses being linked to outcroppings along the Spanish and French coast, from which they parted some 20 million years ago.

Physical features. *Geology.* The majority of geologic units of the Apennines are made up of marine sedimentary rocks that were deposited over the southern margin of the Tethys Sea, the large ocean that spread out between the Paleo-European and the Paleo-African plates during their separation in the Mesozoic era (about 245 to 66 million years ago). These rocks are mostly shales, sandstones, and limestones, while igneous rocks (such as the ophiolites of the northern Apennines, the remains of an older oceanic crust) are scarce. The oldest rocks—metamorphic units of the late Paleozoic era (about 300 to 245 million years ago), with their continental sedimentary cover containing plant remains—represent the relicts of the ancient continental crust of Gondwanaland and are found in small outcroppings. The granitic intrusions and metamorphic units of the Calabrian and Sicilian ranges are also Paleozoic (Hercynian orogeny), but they are believed to be Alpine in origin and only became part of the Apennine chain through subsequent major tectonic movements.

Apennine orogeny

The Apennine orogeny developed through several tectonic phases, mostly during the Cenozoic era (*i.e.,* since about 66 million years ago), and came to a climax in the Miocene and Pliocene epochs (23.7 to 1.6 million years ago). The Apennines consist of a thrust-belt structure with three basic trending motions: toward the Adriatic Sea (the northern and central ranges), the Ionian Sea (Calabrian Apennines), and Africa (Sicilian Range). During Plio-Pleistocene times, ingression and regression of the sea caused the formation of large marine and continental sedimentary belts (sands, clays, and conglomerates) along the slopes of the new chain. In the past million years numerous large faults have developed along the western side of the Apennines, which may be connected to the crustal thinning that began about 10 million years ago and resulted in the formation of a new sea, the Tyrrhenian. Most of these faults have also facilitated strong volcanic activity, and a volcanic chain has formed along them from Mount Amiata in Tuscany to Mount Etna in Sicily; most of these volcanoes—including Mount Amiata, Mount Cimino, the Alban Hills near Rome, and the Ponza Islands—are extinct, but, to the south, Mount Vesuvius, the Eolie Islands, and Mount Etna are all still active. Seismic activity is common along the entire length of the chain (including Sicily), with more than 40,000 recorded events since AD 1000. Mostly earthquakes are shallow (three to 19 miles deep), and their occurrence is probably connected to the settlement of the chain in the complicated interaction between the African and European tectonic plates.

The geologic youth of the Apennines, and a great variety of rock types, are responsible for the rugged appearance of the range today. In the north, in Liguria, sandstones, marls, and greenstones occur. Landslides often occur in these brittle rocks. In Tuscany, Emilia, Marche, and Umbria, clay, sand and limestones are common. In Lazio, Campania, Puglia, Calabria, and northern and eastern Sicily, there are large calcareous rock outcrops, separated by lowland areas of shale and sandstone. In Molise, Basilicata, and Sicily, extensive argillaceous (clayey) rock types occur. Here, the landscape has a thirsty and desolate appearance, with frequent erosion of the *calanchi,* or badlands, type.

Physiography. Starting from the north, the main subdivisions of the Apennines are the Tuscan-Emilian Apennines, with a maximum height of 7,103 feet at Mount Cimone; the Umbrian-Marchigian Apennines, with their maximum elevation (8,130 feet) at Mount Vettore; the Abruzzi Apennines, 9,554 feet at Mount Corno; the Campanian Apennines, 7,352 feet at Mount Meta; the Lucanian Apennines, 7,438 feet at Mount Pollino; the Calabrian Apennines, 6,414 feet at Mount Alto; and, finally, the Sicilian Range, 10,902 feet at Mount Etna. The ranges in Puglia (the "bootheel" of the peninsula) and southeastern Sicily are formed by low, horizontal limestone plateaus, which remained less affected by the Alpine orogeny.

The rivers of the Apennines have short courses. The two principal rivers are the Tiber (252 miles long), which follows a southerly course along the western base of the Umbrian-Marchigian range before flowing through Rome to the Tyrrhenian Sea, and the Arno (155 miles), which flows westerly from the Tuscan-Emilian range through Florence to the Ligurian Sea. In spite of the limited length of the rivers, the action of running water is the chief agent of erosion responsible for molding the contemporary Apennine landscape. The character of the physical geography depends on the varying nature of the rocks in each region and their resistance to water action. The overall aspect of relief, however, exhibits characteristics of an early, or juvenile, stage in the cycle of erosion. In limestone areas, karst erosion, with crevasses worn by water action, predominates. In the highest part of the Apennines there are traces of the erosive action of the glaciers of the last Ice Age, although, unlike the Alps, contemporary glaciers are lacking.

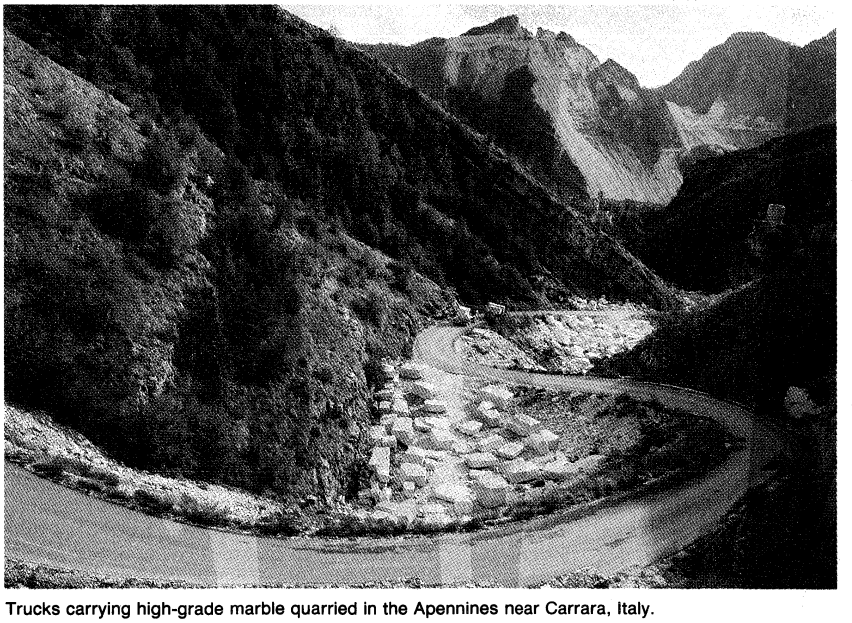
Landscape formed by erosion

Lakes—which today are small and scattered in distribution—were also much more abundant in earlier Quaternary times. The alluvial Lake Trasimeno (49 square miles [128 square kilometres]) in the Umbrian-Marchigian Apennines is the largest lake of the present range. Other natural lakes, of varying origin, are scattered throughout the range. There are more than 200 artificial lakes created for purposes of power and irrigation.

Climate. The climate of the highest section of the Apennines is continental (as found in the interior of Europe) but ameliorated by Mediterranean influences. Snowfalls are frequent, with cold winters and hot summers (average July temperature 75°–95° F [24°–35° C]). Average rainfall—at between 40 and 80 inches (1,000 and 2,000 millimetres) per year—is higher on the Tyrrhenian slopes than on the eastern, or Adriatic, side of the Apennines.

Plant and animal life. The flora of the Apennines is Mediterranean in type and varies with both latitude and altitude. In the north, woodlands with oak, beech, chestnut, and pine predominate. To the south, ilexes, bays, lentisks, myrtles, and oleander (a flowery evergreen herb) abound. Prevailing crops are represented by the olive trees, growing to a height of about 1,300 feet above sea level; citrus fruits, which are well developed in Calabria and Sicily; and grapes, which are found in abundance in Tuscany, Lazio, and Puglia. Other products of the range include sugar beets (in the plain of Emilia), potatoes, vegetables, and fruit. The importance of corn (maize) diminished with the depopulation of hill farms. In the highland areas, pasturing remains the main form of land utilization.

© Kelly W. Culpepper—Photo Researchers



Trucks carrying high-grade marble quarried in the Apennines near Carrara, Italy.

Apennine fauna has been little studied. In addition to typical Mediterranean fauna, many of the indigenous Apennine species (with several species found exclusively within the range, including some insects, the brown "marsicano" bear, the chamois, the wolf, and the wild boar) are now preserved in two natural reserves (Abruzzo National Park and Sila Park) and several regional parks.

Settlement patterns

The people and economy. Since prehistoric times the Apennines have been the home of Italic peoples. Today, the highest village settlement is found at about 4,500 to 5,000 feet above sea level, at the upper limit of cultivated land. More densely populated areas are found in the wide river valleys, which are rich in alluvial and cultivated land (*e.g.*, the valleys of Lunigiana in Liguria, Garfagnana in Tuscany, and those of the upper Arno and Tiber rivers). Internal basins (Foligno, Terni, Rieti, l'Aquila, Sulmona, Avezzano) are also well populated. Rural depopulation, resulting from the lack of development of the Italian south and the attraction of industrial areas in northern Italy and elsewhere in Europe, has reached major proportions. This emigration has nevertheless slackened, mainly as a result of attempts to develop the local economy.

In the foothills of the Apennines, manufacturing industries are widespread, while extraction industries have been developed in the adjacent coastal plain, often in association with important discoveries of natural gas. Such minerals as mercury, sulfur, boron, and potassic salts are also of significance, while the marble quarries—particularly those near Carrara—of the Apennines have been famous for centuries.

The Apennines are crossed by several railway lines, some of them double-tracked. There are numerous roads providing access to the range, although the rugged terrain makes for difficulties. Among the highways that have overcome the barriers of relief with imposing series of tunnels and embankments is the Autostrada del Sole ("Highway of the Sun"), which is the main artery of peninsular Italy and one of the great scenic routes of Europe.

Study and exploration. Various aspects of the Apennines—their geology, hydrography, zoology, and botany—have been studied by the leading Italian universities, the Italian Geological Survey, and such bodies as the National Research Council of Italy and the Hydrographic Service of the Ministry of Public Works. Since the late 1970s many scientists have organized several national research projects concerning the geologic evolution and hazards of the Apennines and have also conducted environmental evaluations and petroleum surveys. (B.A./Ma.P.)

CARPATHIAN MOUNTAINS

The Carpathian Mountains are a geologically young European mountain chain forming the eastward continuation of the Alps. From the Danube Gap, near Bratislava, Czech., they swing in a wide arc some 900 miles (1,450 kilometres) long to near Orşova, Rom., at the portion of the Danube River valley called the Iron Gate. These are the conventional boundaries of these arcuate ranges, although, in fact, certain structural units of the Carpathians extend southward across the Danube at both sites mentioned. The true geologic limits of the Carpathians are, in the west, the Vienna Basin and the structural hollow of the Leitha Gate in Austria and, to the south, the structural depression of the Timok River in Yugoslavia. To the northwest, north, northeast, and south the geologic structures of the Carpathians are surrounded by the sub-Carpathian structural depression separating the range from other basic geologic elements of Europe, such as the old Bohemian Massif and the Russian, or East European, Platform. Within the arc formed by the Carpathians are found the depressed Pannonian Basin, composed of the Little and the Great Alfolds of Hungary, and also the relatively lower mountain-and-hill zone of Transdanubia, which separates these two plains. Thus defined, the Carpathians cover some 80,000 square miles (200,000 square kilometres).

Although a counterpart of the Alps, the Carpathians differ considerably from them. Their structure is less compact, and they are split up into a number of mountain blocks separated by basins. The highest peaks, Gerlachovský Štít (Gerlach) in the Carpathians (8,711 feet [2,655 metres]) and Mont Blanc in the Alps (15,771 feet), differ greatly in altitude, and in average elevation the Carpathian mountain chains are also very much lower than those of the Alps. Structural elements also differ. The sandstone–shale band known as flysch, which flanks the northern margin of the Alps in a narrow strip, widens considerably in the Carpathians, forming the main component of their outer zone, whereas the limestone rocks that form a wide band in the Alps are of secondary importance in the Carpathians. On the other hand, crystalline and metamorphic (heat-altered) rocks, which represent powerfully developed chains in the central part of the Alps, appear in the Carpathians as isolated blocks of smaller size surrounded by depressed areas. In addition to these features, the Carpathians contain a rugged chain of volcanic rocks.

Similar differences can be observed in the relief of these two mountain systems, notably in the way that the processes of erosion have occurred. The relief forms of the Alps today result for the most part from the glaciations of the last Ice Age. These affected practically all mountain valleys and gave them their specific relief character. In the Carpathians, glaciation affected only the highest peaks, and the relief forms of today have been shaped by the action of running water.

Physical features. *Geology.* The Carpathians extend in a geologic system of parallel structural ranges. The Outer Carpathians—whose rocks are composed of flysch—run from near Vienna, through Moravia, along the Polish-Czechoslovak frontier, and through the western Ukraine into Romania, ending in an abrupt bend of the Carpathian arc north of Bucharest. In this segment of the mountains, a number of large structural units of nappe character (vast masses of rock thrust and folded over each other) may be distinguished. In the eastern part of the Outer Carpathians this fringe is formed by the Skole Nappe, and in the western part it is formed by the Silesian Nappe, both of which are split by the longitudinal central Carpathian depression. Overthrust on the Silesian Nappe is the Magura Nappe, the counterparts of which in the east are the Chernogora (Chornohora) and the Tarcău nappes.

Parallel structural ranges

The Inner Carpathians consist of a number of separate blocks. In the west lies the Central Slovakian Block; in the southeast lie the East Carpathian Block and the South Carpathian Block, including the Banat and the East Serbian Block. The isolated Bihor Massif, in the Apuseni Mountains of Romania, occupies the centre of the Carpathian arc. Among the formations building these blocks are ancient crystalline and metamorphic cores onto which younger sedimentary rocks—for the most part limestones and dolomites of the Mesozoic era (245 to 66.4 million years ago)—have been overthrust.

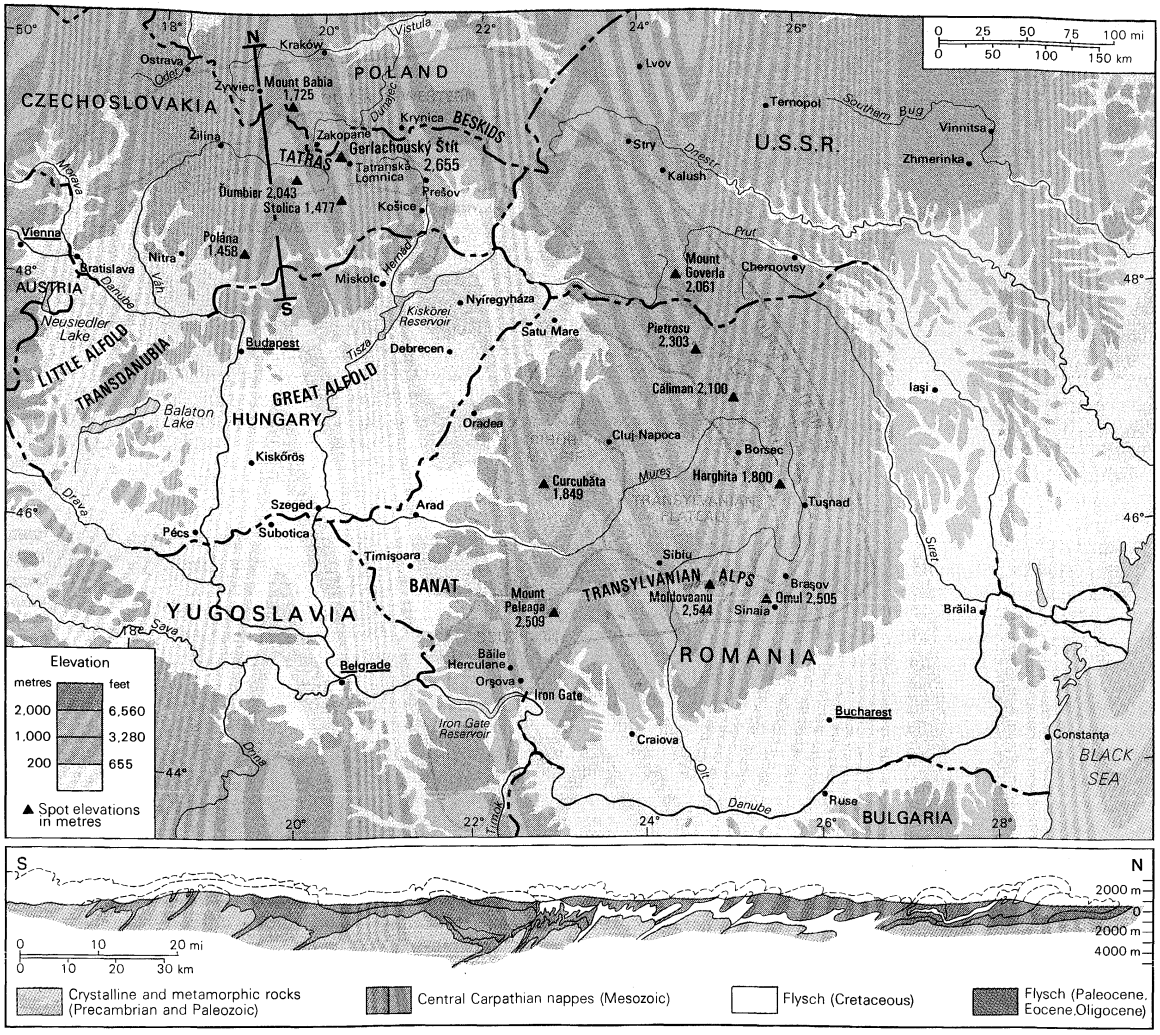
The third and innermost range is built of young Tertiary volcanic rocks formed less than 50 million years ago, differing in extent in the western and eastern sections of the Carpathians. In the former they extend in the shape of an arc enclosing, to the south and east, the Central Slovakian Block; in the latter they run in a practically straight line from northwest to southeast, following the line of a tectonic dislocation, or zone of shattering in the Earth's crust, parallel with this part of the mountains. Between this volcanic range and the South Carpathian Block, the Transylvanian Plateau spreads out, filled with loose rock formations of young Tertiary age.

The Central Slovakian Block is dismembered by a number of minor basins into separate mountain groups built of older rocks, whereas the basins have been filled with younger Tertiary rocks.

In Romania, orogenic, or mountain-building, movements took place along the outer flank of the Carpathians until late in the Tertiary period (less than 10 million years ago), producing foldings and upheaval of the sedimentary rocks of the sub-Carpathian depression; the result was the formation of a relatively lower range called the sub-Carpathians adjoining the true Carpathians.

Relief forms

The relief forms of the Carpathians have, in the main, developed during young Tertiary times. In the Inner Carpathians, where the folding movements ended in the Late Cretaceous epoch (97.5 to 66.4 million years ago),

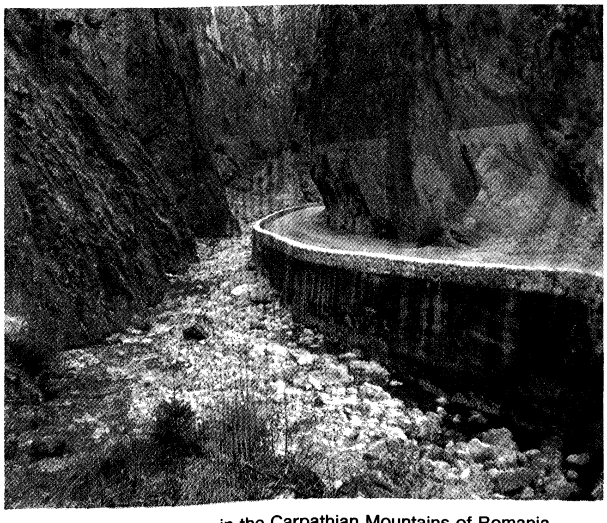


Regional division of the Carpathian Mountains and a geologic cross section of the Western Carpathians. The location of the cross section is shown by the line N–S on the map.

local traces of older Tertiary landforms have survived. Later orogenic movements repeatedly heaved up this folded mountain chain, leaving a legacy of fragmentary flat-topped relief forms situated at different altitudes and deeply incised gap valleys, which often dissect the mountain ranges. In this way, for example, the gap sections of the Danube and of some of its tributaries—the Váh, the Hernád, and the Olt—developed.

The last Ice Age affected only the highest parts of the Carpathians, and glaciers were never more than about 10 miles long, even in the Tatras, where the line of permanent snow ran at 5,500 feet above sea level.

Harrison Forman



Deep river-cut gorge in the Carpathian Mountains of Romania.

Physiography. Generally speaking, the Carpathians have been divided into the Western and the Eastern Carpathians, the latter also called—probably more accurately—the Southeastern Carpathians. The extent of these two regions and their subdivisions is given in Table 1. There are marked differences between these parts. The Western Carpathians show a clearly marked zoning in geologic structure and relief forms, and the highest elevations occur in the central part of this province, in the Tatras and the Lower Tatras ranges. The geologic structure of the inner part of the Western Carpathians is marked by a break running from the east and the south along a line of dislocation in the Earth's crust. Along this line, masses of volcanic rocks have been piled up surrounding the Central Western Carpathian Block in a wide arc, with its convex side turned eastward. The boundary between the Western and the Southeastern Carpathians occurs at the narrowest part of the mountain range, marked by the valley of the San River to the north and the Łupków Pass (2,100 feet) and the Laborec Valley to the south. There the Carpathians are only some 75–80 miles wide, while in the west they are 170 miles and in the east as much as 220–250 miles across.

The Southeastern Carpathians are formed by a triangular block of mountains surrounding a basin. The three mountain formations concerned differ in origin and structure. The Eastern Carpathians, running in a northwest–southeast direction, include the flysch band, which represents the continuation of the Outer Western Carpathians, and also an inner band of crystalline and volcanic rocks. In

Table 1: Subdivisions of the Carpathians

	approximate area	
	sq km	sq mi
Western Carpathians	68,000	26,000
Outer Western Carpathians	27,500	10,500
Central Western Carpathians	15,500	6,000
Inner Western Carpathians	25,000	9,500
Southern Carpathians	131,000	50,500
Outer Eastern Carpathians	35,500	13,500
Inner Eastern Carpathians	21,250	8,000
Southern Carpathians	28,250	11,000
Transylvanian Plateau	28,500	11,000
Bihor Massif (Apuseni Mountains)	17,500	7,000

contrast, the Southern Carpathians, running east-northeast to west-southwest, consist, in the main, of metamorphic rocks. The Bihor Massif is also of metamorphic rock but is covered with younger sediments.

The Outer Western Carpathians are generally of low altitude; the highest elevation is Mount Babia (5,659 feet) in the Beskid Range, straddling the borders of Poland and Czechoslovakia. On the Polish side, a national park has been established. A considerable part of the Outer Western Carpathians lacks a truly mountainous landscape and rather resembles a hilly plateau elevated to 1,300–1,600 feet above sea level.

The Tatras

The Central Western Carpathians consist of a series of isolated mountain ranges separated by structural depressions. Highest among them are the Tatras (Gerlachovský Štít, 8,711 feet), exhibiting a typical high-mountain glacial relief with ice-scoured (cirque) lakes and waterfalls. This highest Carpathian massif is built of crystalline (granite) and metamorphic rocks, but the northern part contains, upthrust from the south, several series of limestone rocks with associated karst, or water-incised, relief forms. On both the Polish and Slovakian sides, national parks have been established. South of the Tatras, separated by the Liptov and Spiš basins, run the parallel Lower Tatras, similar in geologic structure but lower (Ďumbier Peak, 6,703 feet) and with a less conspicuous glacial relief. Along the boundary line between the Outer and the Central Western Carpathians extends a narrow strip of klippen (limestone) rocks, which, north of the Tatras, has developed into the small but picturesque Pieniny mountain group. A narrow and sharply winding gap valley has been incised there by the Dunajec River, a tributary of the Vistula.

The Inner Western Carpathians are lower and more broken. The principal mountain groups are the Slovak Ore Mountains (Slovenské Rudohorie), with Stolica (4,846 feet) as the highest peak; they are built of metamorphic rocks and of sedimentaries of the Paleozoic era more than 250 million years old. Also found there are tableland areas of Mesozoic limestones, about 150 million years old, containing such large caves as the Domica-Aggtelek Cave on the Slovak-Hungarian boundary, which is 13 miles long. Mountain groups of volcanic origin are important in this part of the Carpathians; the largest among them is Pol'ana (4,784 feet).

Compared with the Outer Western Carpathians, the Outer Eastern Carpathians, which are their continuation, are higher and show a more compact banded structure. The highest mountain group is the Chernogora on the Ukrainian side, with Goverla (Hoverla; 6,762 feet) as the highest peak. The Inner Eastern Carpathians attain their highest altitude in the Rodna (Rodnei) Massif in Romania; they are built of crystalline rocks and reach a peak in Pietrosu (7,556 feet). To the south, extinct volcanoes in the Căliman and Harghita ranges have, to some extent, kept their original conical shape; the highest peaks of these ranges are 6,890 feet and 5,906 feet, respectively. Fringing the true Eastern Carpathians runs a narrow zone called the sub-Carpathians, which is made up of folded young Tertiary rocks superimposed on the sub-Carpathian structural depression.

The Southern Carpathians culminate in the Făgăraş Mountains (highest point Moldoveanu, 8,347 feet), which show Alpine-type relief forms. The western part of the Southern Carpathians—that is, the Banat Mountains and the mountains of eastern Serbia (which, at the Iron Gate, are split apart by the gap valley of the Danube)—do not exceed an altitude of 5,000 feet.

The Bihor Massif, which occupies an isolated position inside the Carpathian arc, features widespread flat summit plains bordered by narrow, deep-cut valleys. The highest peak is Curcubăta (6,067 feet).

Finally, mention should be made of the Transylvanian Plateau. This is made up of poorly resistant young Tertiary rocks and characterized by a forestless hilly landscape with elevations of 1,500 to 2,300 feet above sea level; the valleys are cut to depths of 325 and 650 feet.

Drainage. The water runoff from the Carpathians escapes for the most part (about 90 percent) into the Black Sea. The great curve of the mountain chain abuts in the south upon the Danube; in the east it is flanked by a tributary of the Danube, the Prut River, and farther on by the Dnestr River, which flows to the Black Sea. Only the northern slope of the Carpathians, mostly in Poland but partly in Czechoslovakia, is linked to the Baltic Sea by the drainage basins of the Vistula and (in part) Oder rivers. Larger rivers originating in the Carpathians include the Vistula and the Dnestr and the following Danube tributaries: Váh, Tisza, Olt, Siret, Prut. The Carpathian rivers are characterized by a rain–snow regime; high-water periods occur in the spring (March–April) and in summer (June–July), with the latter usually more powerful. Often these floods assume catastrophic dimensions caused by the poor ground retention of the rainfall. There has long been an urgent need for the construction of storage basins, work on which was initiated on a large scale in the decades following World War II. The largest storage basin is in the Danube River valley on the frontier between Romania and Yugoslavia. Other large basins include one in the Bistriţa valley in Romania, one in the San valley in Poland, and one in the Orava valley in Czechoslovakia. Altogether there are some 50 storage basins in the Carpathians. Natural mountain lakes are relatively rare, and all of them are small. Although there are some 450 lakes, their total surface is barely 1.5 square miles. The high-mountain lakes are mainly of glacial origin.

Flood peril

Climate. The situation of the Carpathians, on the boundary line between western and eastern Europe, is reflected in the features of their climate, which in winter is governed by the inflow of polar-continental air masses arriving from the east and northeast, while during other seasons oceanic air masses from the west predominate. The distance from the Atlantic Ocean (from 620 to 1,240 miles) and the influence of the intervening masses of the Alps and the Bohemian Massif cause diminished precipitation in the Carpathians. The Carpathians thus possess certain features of a continental climate, although from the viewpoint of relief they constitute a sort of island amid the surrounding plains, where the climate is much drier. The continentality of the climate is clearly seen in the intermontane depressions, however, as well as on the lower parts of the southern mountain slopes. In winter, temperature inversion, in which the low depressions retain very cold air while the mountaintops show relatively high temperatures, is a common occurrence throughout the Carpathians. In some depressed areas, notably the Transylvanian Plateau, the total annual precipitation is less than 24 inches (600 millimetres), while precipitation in the mountains at 2,600 feet (800 metres) above sea

Table 2: Climatic Stages of the Western Carpathians

type	stage	mean annual temperature		average altitude limits (above sea level)*	
		degrees Fahrenheit	degrees Celsius	feet	metres
Nival	cold	25	−4	8,710	2,655
Nival-pluvial	temperate cold	28	−2	6,070	1,850 (1,670)
	very cool	32	0	5,080	1,550 (1,400)
	cool	36	2	3,600	1,100
Pluvial-nival	temperate cool	39	4	2,300	700
	temperate warm	43	6	820	250
	mountain foreland	46	8	under 820	under 250

*The figures in parentheses refer to the Outer Carpathians.

Table 3: Vegetation Stages of the Carpathians
(feet [metres])

stages	Western Carpathians		Eastern Carpathians	Southern Carpathians
	Outer	Inner		
Nival	—	up to 8,710 (2,655)	—	—
Alpine	up to 5,660 (1,725)	up to 7,200 (2,200)	up to 6,600 (2,000)	up to 8,344 (2,544)
Subalpine	up to 5,480 (1,670)	up to 5,900 (1,800)	up to 6,070 (1,850)	up to 7,200 (2,200)
Upper forest	up to 4,600 (1,400)	up to 5,080 (1,550)	up to 5,080 (1,550)	up to 5,900 (1,800)
Lower forest	up to 3,770 (1,150)	up to 4,100 (1,250)	up to 4,100 (1,250)	up to 4,900 (1,500)
Foreland	up to 1,800 (550)	up to 2,300 (700)	up to 2,000 (600)	up to 2,800 (850)

level is about 45 inches, and on the highest massifs it reaches 65 to 70 inches. The mean annual and monthly air temperatures vary according to altitude above sea level but by no means at constant rates.

For the Polish part of the Carpathians, a series of climatic types and stages has been distinguished; and with slight modification these may be applied to the whole Carpathian mountain range.

Plant and animal life. Different vegetation stages may also be distinguished for the various altitudinal zones of the Carpathians. The alpine stage is characterized by high mountain pastures, the subalpine stage by dwarf pine growth, the upper forest stage by spruce, and the lower forest stage by beech. The foreland stage is noted for oaks and elms. The natural vegetation stages are matched by stages of economic land use: the foreland by wheat and potato growing, the lower forest stage by oats and potato growing (up to 3,280 feet), and the upper forest stage and the subalpine stages by pastoral use.

The plant life of the Carpathians contains many unique species, especially in the southeastern part of the mountains where the effect of Quaternary climatic cooling was less marked. Forests have been best preserved in the eastern part of the Carpathians, and there the animal life includes bears, wolves, lynx, deer, boars, and, in the highest parts (in the Tatras), chamois and marmots.

The people. The distribution of the population in the Carpathians depends on natural land features and on socioeconomic conditions; hence it is very much diversified. In the valleys between the mountains and again on the northern slopes of the Western Carpathians, the population density is heavy, whereas close by practically uninhabited mountain massifs are to be found. On the whole, the Southeastern Carpathians are less densely settled than the Western Carpathians, but there also marked aggregations of people occur in the valleys.

Nationalities

The western slope of the Western Carpathians is inhabited by Czechs, the northern slope by Poles, the entire central part of the Western Carpathians by Slovaks, and the southern portion by Hungarians. The northern part of the Eastern Carpathians, both its outer and inner sectors, is occupied by Ukrainians; but south of latitude 47° a Romanian population predominates. Inside the arc of the Eastern Carpathians, and also partly on the Transylvanian Plateau, lives a compact island of Hungarian population and some remnants of German colonists dating from the Middle Ages. Finally, the southwestern margin of the Carpathians, beyond the Danube gap, is occupied by Serbs. Generally speaking, the greater part of the Western Carpathians and the northern part of the Eastern Carpathians is inhabited by a Slav population, and the southern part of both these Carpathian provinces, with the exception of the mountains of east Serbia, by Romanians and Hungarians.

In the 13th and 14th centuries Romanian shepherds, wandering with their flocks, moved along the Carpathians into what is today Ukrainian, Slovakian, and Polish territory, and traces of this penetration have survived in geographic nomenclature and in economic methods and also in types of buildings, garments, and customs, although by the second half of the 20th century many of the latter were gradually disappearing. In general outlines, but by no means in detail, the diversity in nationality coincides with today's pattern of the political boundaries.

The economy. *Agriculture and industry.* The Carpathians are a region of agriculture and forestry, with industry in an early stage of development. Agriculture flourishes on the Transylvanian Plateau, in intramontane basins, and on lower parts of the mountains, up to some 3,000 feet elevation. On the northern slopes wheat, rye, oats, and potatoes predominate; on the southern slopes corn (maize), sugar beets, grapes, and tobacco are grown. Above 3,000 feet elevation forestry and pastoral life are the rule. Natural gas, found mainly on the Transylvanian Plateau, is important among natural resources. Oil is also significant; the richest deposits lie in the Romanian sub-Carpathians. Brown coal is found in low-lying areas of the Western Carpathians in Czechoslovakia and Hungary, and some bituminous coal is mined in the Romanian Southern Carpathians. Also noteworthy are the rock salt beds of the Transylvanian Plateau, the Romanian sub-Carpathians, and the base of the Polish Carpathians, and the potassium salts found at the base of the Ukrainian Carpathians. Iron ores, ores of noniron metals, and gold and silver ores were intensively mined in the Middle Ages in the Bihor Massif and in the Slovakian Western Carpathians, but today all these deposits are of minor importance.

Economic development

Larger industrial centres are Bratislava, the capital of the Slovak Republic, with a thriving machinery and a petrochemical industry; and Košice, the principal town of eastern Slovakia, with a modern steel mill. Prominent in Romania are Cluj-Napoca, which is the principal town of the Transylvanian Plateau, concentrating on machinery making and chemical and food products; Braşov, situated in a basin near the boundary between the Western and Southern Carpathians, a town where machine production predominates; and Sibiu, lying between the Transylvanian Plateau and the Southern Carpathians.

Tourism. The Carpathians are a popular tourist and recreation venue, especially for the people of Poland, Czechoslovakia, Hungary, and Romania. Tourist travel from other countries is less developed, although a number of areas attract visitors from abroad. Most important among these are Zakopane, a centre of sports activities, tourism, and recreation, situated in Poland north of the Tatras. On the Slovak side of the Tatras, a similar role is played by a number of localities, notably Tatranská Lomnica, Smokovec, and Štrbské Pleso. In Romania the outstanding centre for winter sports and tourism is Sinaia, situated in the Prahova valley. The Carpathians are noted for their abundance of mineral springs. Among the best-known Carpathian health spas are Krynica in Poland, Piešťany in Czechoslovakia, and Borsec, Băile Herculane, and Tuşnad in Romania.

Transportation. The railway network of the Carpathians came into existence in the latter half of the 19th century and the beginning of the 20th, at a time when most of the mountains were in Austria-Hungary. In this period the nodal point was Budapest, situated in the centre of the Carpathian arc. The principal railway lines were laid out radially from Budapest across the various mountain passes and were tied in with the main longitudinal west–east trunk line running in an arc along the northern flank of the Carpathians between Vienna and Chernovtsy, Ukrainian S.S.R. (then situated in Austria-Hungary). This northern trunk continued in the sub-Carpathian Romanian railway line running toward Bucharest and, farther on, to Orşova, which, in turn, was linked by a Hungarian railway section with Budapest and thus with Vienna. After the Austro-Hungarian Empire had collapsed, this system lost much of its economic and strategic importance. Within its boundaries the new state of Czechoslovakia started to build

The historical legacy

longitudinal west–east railway lines. For Romania, which had been allotted Carpathian Transylvania, the previously neglected lines became highly important. To some extent, a change in this pattern came about after World War II, when the northern part of the Eastern Carpathians and Trans-Carpathian Ukraine became part of the Soviet Union. The railway lines crossing this part of the Carpathians became arteries linking the Soviet Union, Czechoslovakia, and Hungary. Although the lines between Poland and Slovakia lost most of their importance in passenger and freight transport, truck routes utilizing the Dukla (1,640 feet), Jablonkov, and other passes became significant in freight traffic between Poland and the countries south of the Carpathians. The most important Carpathian railway lines have been electrified, although the Budapest–Vienna line was electrified before World War II.

Study and exploration. Many nationalities are in contact with one another in the Carpathians, and this diversity has had its effect on the development of scientific research in the region. From the end of the 18th century until World War I, most of the Carpathians were within the boundaries of Austria-Hungary, and throughout this period the Carpathians were readily accessible to all scientists of this multinational empire; the work of Polish scientists, together with that of Germans and Hungarians, is considered most noteworthy. In the late 19th century the

Mapping

Austrian general staff published the first comprehensive topographic map of the region. A century later, each of the countries whose territory covered part of the Carpathians—Czechoslovakia, Poland, Romania, Hungary, and the Soviet Union—had at hand topographic maps drawn to a scale of 1:50,000 and 1:200,000—compiled on the basis of a coordinated geodetic system and in a mutually correlated sheet pattern.

As for geologic maps, the first paper dealing with the geology of the Carpathians as a whole was published in 1815. Today, each of the Carpathian countries has its own general geologic maps, and there is also abundant regional geologic literature. In 1922 the International Geological Congress created an association of Carpathian geologists, which met every three years thereafter. Regional research in physical geography is also well advanced, and in 1963 a geomorphologic committee for the Carpathians and the Balkans was established.

Research is somewhat less advanced in climatology and biogeography, although a number of papers began to appear in the second half of the 20th century. In human geography much attention has been given to the problems of pastoral life and associated population movements. No synthetic survey of the economic geography of the whole Carpathians has appeared as economic problems have been studied separately in each of the countries involved. Indeed, the first comprehensive geographic account of the Carpathians as a whole, by the Polish geographer Antoni Rehman, was not published until 1895.

Since World War II the Carpathians have become the object of research by a number of scientific centres in the countries involved, with the geographic institutes of the several national academies of sciences and the geographic and natural history institutes of various universities playing a leading role. National geologic institutes and institutes of hydrology and meteorology have also amassed a considerable body of information. (J.A.K.)

EUROPEAN PLAIN

One of the greatest uninterrupted expanses of plain on the Earth's surface sweeps from the Pyrenees Mountains on the French-Spanish border across northern Europe to the Ural Mountains in the Soviet Union. In western Europe the plain is comparatively narrow, rarely exceeding 200 miles (320 kilometres) in width, but as it stretches eastward it broadens steadily until it reaches its greatest width in the Soviet Union, where it extends more than 2,000 miles.

Because it covers so much territory, the plain gives Europe the lowest average elevation of any continent. The flatness of this enormous lowland, however, is broken by hills, particularly in the west.

Physical features. *Physiography.* The western and central European section of the plain covers all of western and northern France, Belgium, The Netherlands, southern Scandinavia, northern Germany, and nearly all of Poland. From northern France and Belgium eastward it is commonly called the North European Plain.

The North European Plain

Conditions in the North European Plain are complex in detail. The terrain is flat or gently undulating. Most of the area was glaciated several times during the Pleistocene epoch (1.6 million to 10,000 years ago), and the landscape is typically postglacial. Drainage is poorly developed, glacial deposits called moraine blanket much of the area, and large sections are underlain by glacial outwash plains. Hilly terminal moraines, marking the stationary edges of the Pleistocene ice sheets, are strewn in great arcs across northern Germany, Poland, and the European sections of the Soviet Union. Interspersed with these moraines are long parallel spillways where glacial meltwaters flowed to the sea parallel to the ice front. These spillways were covered with sand and gravel by the rushing glacial streams. Today they are occupied by flat, poorly drained wetlands that are relatively unproductive. Sandy duneland borders the North and Baltic seas, and extensive windblown loess deposits, resulting from the intense wind erosion of the barren interglacial and postglacial landscapes, stretch across the North European Plain from France to the Soviet Union.

Other landforms in the North European Plain include the extensive delta plain of The Netherlands that is formed by the deposits of the Rhine River as it enters the North Sea. Like many other delta plains, this area has rich and fertile soils and a flat terrain that is favourable for agriculture where it has been properly drained. The Rhine has historically provided excellent transportation, and the region is one of the most densely populated areas in the world.

The East European Plain

In the east the plain is called the East European Plain (Russian: Vostochno-Yevropeyskaya Ravnina). Finland in the northwest is underlain by ancient, resistant, crystalline rocks, part of the Precambrian Scandinavian Shield. Because it was near the origin of the Pleistocene ice sheets that advanced southward over continental Europe, Finland's landscape is characterized more by glacial erosion than by glacial deposition. With its numerous lakes and swamps caused by the disarranged and immature drainage pattern, together with its thin soils and coniferous forests, the Finnish plain is similar in character and appearance to northern and eastern Canada, another heavily glaciated Precambrian Shield area. The continental glaciers that planed, eroded, and polished the rock surfaces in Finland deposited part of the material over the plains to the south.

The European section of the Soviet Union is deeply underlain by a relatively rigid platform of ancient rocks. At various times in its history, however, this area has slipped beneath the sea and been covered with sedimentary rocks. These rocks have been mildly bent and warped, but nowhere have they been sharply deformed. Consequently, the whole area from the Black Sea to the Arctic is one uninterrupted plain, everywhere below 1,500 feet (450 metres) in elevation.

Climate. The climate on the whole is characterized by marked seasonal changes, with cold winters and warm summers. The west has a maritime climate very favourable to agriculture. It has enough rain in all seasons to keep fields green. Summers are warm but not hot, and winters are cold but not freezing. As one moves eastward, the ameliorating maritime influence diminishes, and the character of the climate becomes more continental: rainfall is concentrated in the warmer months, summers are hotter, and winters become extremely cold. Spring and fall nearly disappear as separate seasons, and the greenness of the summer gives place abruptly each year to the gray drabness of a frozen winter. Agriculture in eastern Europe tends to be more difficult and less productive than in the west.

Drainage. The Garonne and the Loire rivers, with their numerous tributaries, drain much of western France before they enter the Bay of Biscay, and the Seine crosses the broad synclinal lowland of the Paris Basin on its way to the English Channel. The Scheldt and its affluents (Lys, Scarpe, Dender, Demer) drain the Plain of Flanders and the low plateaus of central Belgium. The Meuse pursues a

Rolling expanse of the East European Plain, consisting of glacial deposits, in southern Poland.
D.C. Williamson

varied course through the scarplands of Lorraine, crosses the Ardennes in a valley cut transversely to the structure, turns at right angles along the coal furrow of southern Belgium, and then in a sweeping curve flows across the plain of the southern Netherlands to form a joint floodplain with the lower Rhine.

The Rhine is the main river of west central Europe, 865 miles in length, crossing the various structural and relief zones from its Alpine sources and entering its plain course in the North Sea lowlands. Farther east the several broadly parallel systems include the Weser, the Elbe, the Oder, and the Vistula, which rise in the uplands of central Europe and flow in a general northwesterly direction across the lowlands to the North or Baltic Sea. Each of these rivers reveals distinct right-angle bends, the result of the Pleistocene ice sheets, the margins and terminal moraines of which lay along an east–west line so that meltwaters escaping to the sea had to flow in a westerly direction, eroding broad intermorainal channels (*Urstromtäler*). When the ice sheets withdrew, the rivers occupied some sections of the east–west meltwater channels between their northerly courses. During the 19th and early 20th centuries several of these channels were used as routes to construct canals linking the north-flowing rivers. The Mittelland Canal between East and West Germany is the most prominent of these.

Plant and animal life. Deciduous and coniferous forests diversify the landscape of the North European Plain, although present forests are no more than remnants of a thick mixed forest of oak, elm, ash, linden, and maple, which, since the Middle Ages, has given way to villages and fields in most places. The East European Plain, despite its great uniformity in terrain, exhibits strong regional contrasts in vegetation. Climatic differences produce great belts of characteristic plant life extending approximately east–west across the country. The southern part of the plain is an area of semiarid grasslands, which grade toward the north into more humid lands with taller grasses and rich, fertile soils. North of the grasslands lies a belt of hardwood forests; in the severely cold north lies a belt of coniferous forests and, bordering the Arctic Ocean, a belt of tundra.

Forest remnants

The wild animals of the plain are those characteristic of the whole of Europe, but their numbers have been considerably reduced and their habitats modified by intense human settlement of most areas of the plain.

The people. A variety of languages is spoken on the plain. As in the rest of Europe, almost all belong to the Indo-European family. The primary exceptions are the Finno-Ugric languages Hungarian and Finnish. Three major branches of Indo-European speech are represented. The Germanic branch is represented by Dutch, Flemish (in part of Belgium), German (including the dialect of Austria), Danish, and Swedish. The major representative of the Romance branch is French, along with some inhabitants of the East European Plain who speak Romanian. In most of the east, however, people speak languages of the Slavic branch, of which Polish, Russian, Belorussian (White Russian), and Ukrainian are the most widespread, but also including Czech, Slovak, Slovene, Serbo-Croatian, and Bulgarian.

The European Plain includes people as diverse in culture as the French, Russians, Hungarians, and Swedes. In spite of cultural differences, many of these peoples traditionally shared underlying similarities that derived in part from a common pattern of village life and agricultural routine. Although the eastern part of the plain has remained traditional in many ways, the western part has been transformed as urban centres and industrialization have expanded into the surrounding countryside. Modern transportation has made it easy for farmers to get regularly to town. In many places, members of farming families, released from working on the land by the efficiency of modern machinery, find jobs in town but without moving from the farm. Conversely, urbanites find that they can live in villages and work in town. Where this has occurred, the centuries-old distinction between urban and rural cultures (or subcultures) has been obliterated; even where developments have not gone that far, to the extent that the farmer is no longer parochial, the old distinction has been broken down. (R.T.A./Ed.)

The economy. The European Plain has long been a region of major agricultural importance, and, apart from the relatively small area occupied by its cities and towns today, the landscape—especially in the east—remains predominantly agricultural. Since the mid-19th century, however, the plain has also been one of the world's major heavy industrial regions. This has been especially true in the west, where the industrial concentration extending from West Germany's Ruhr valley north along the Rhine River and west into The Netherlands, Belgium, and northern France has become Europe's most important centre of coal, steel, and chemical production. Similar industrial concentrations have grown up around smaller coalfields farther east, notably in East Germany's Westphalia and Poland's Upper Silesia regions and in the Donets coalfield of the Soviet Ukraine. The increasing importance of bulky imported raw materials to Europe's economy since the end of World War II has made large seaports such as Hamburg and Rotterdam major centres of industry and commerce as well.

Heavy industry

History. Parts of the European Plain harboured hunter-gatherer groups through much of the late Pleistocene, but significant settlement on the plain did not begin until postglacial times. Immigrants from the south moved north after the glaciers retreated and settled in widely scattered areas along the seacoasts, rivers, and lakeshores of the then heavily wooded plain. Until about 3000 BC—when agriculture became widespread in northern Europe—hunting, fishing, and foraging with stone and bone tools was the characteristic mode of life on the plain.

The first agricultural settlements were made primarily on lightly wooded sites with porous, easily worked, and well-drained soils—*i.e.*, those most suited to the fragile wooden

tools of the time. Such areas were found on the loess belt of the northern plain, which became the principal region of prehistoric settlement north of the Alps. Settlement on the thickly forested clay soils of the lowlands did not become feasible until the 8th century AD, with the invention of the heavy-wheeled plow. One of the most significant technological inventions of the Middle Ages, the heavy-wheeled plow opened the European Plain to settlement as never before and was soon followed by other improvements in agrarian technology.

The traditional two-field system of crop rotation, in which half the agricultural land was left fallow each year to maintain soil fertility, gave way to the more sophisticated three-field system: in addition to the usual sowing of wheat, barley, or rye in the autumn, another part of the land was planted in oats or nitrogen-fixing legumes (peas and beans) in the spring, and only the remaining third of the land was left fallow. The cultivation of a surplus of oats from the spring planting, moreover, provided feed that made possible the substitution of the swifter-gaited horse for plowing in place of the oxen. The ever-larger agricultural surpluses resulting from these advances led to the establishment of towns—and eventually cities—on the European Plain.

The character of the European Plain, however, remained primarily rural and agricultural until the 19th century, when the Industrial Revolution spread from England onto the Continent. Major coalfields stretching along the North European Plain from the Franco-Belgian border to the Donets Basin in the Soviet Union became focal points for the development of heavy industry. The plain's excellent system of rivers and canals furnished a network for the transport of such bulk cargo as coal and iron ore, and, when faster bulk transport later became necessary, its flat terrain enabled the unhindered construction of an extensive rail system. (Ed.)

The Industrial Revolution

PYRENEES

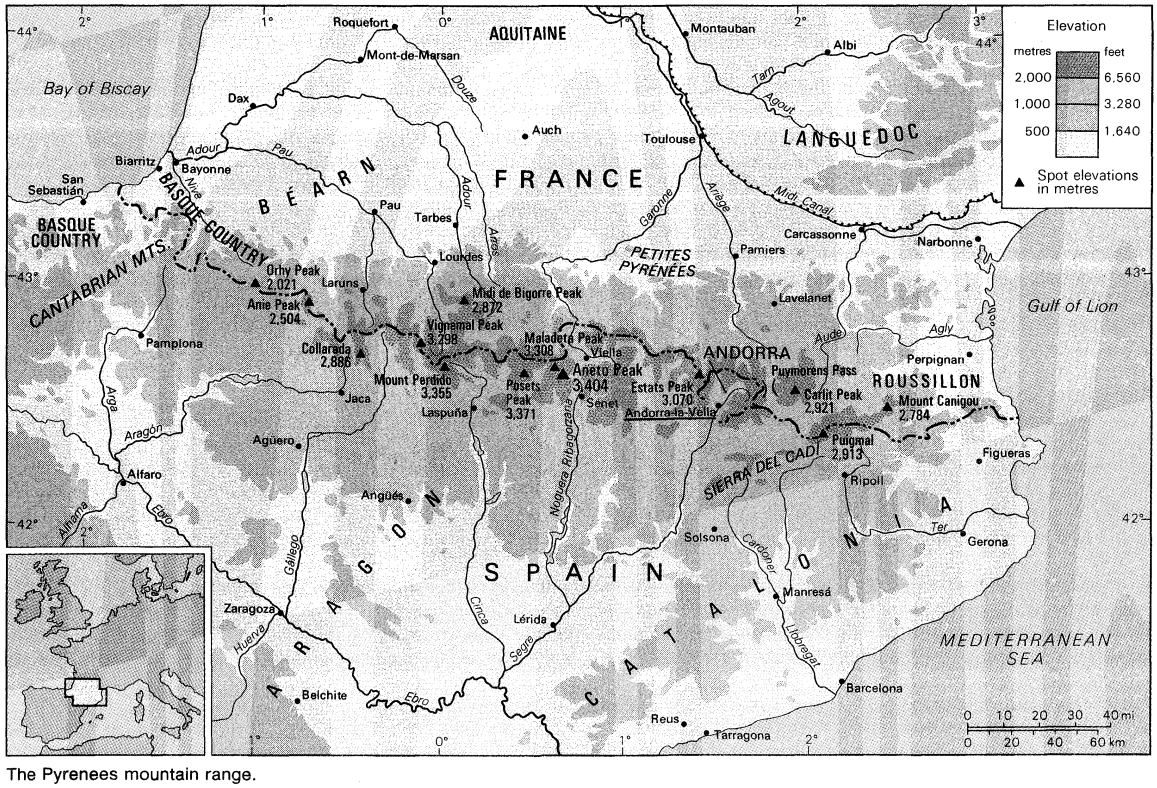
A mountain chain stretching from the shores of the Mediterranean Sea on the east to the Bay of Biscay on the Atlantic Ocean on the west, the Pyrenees (French: Pyrénées; Spanish: Pirineos) form a high wall between France and Spain that has played a significant role in the history of both countries and of Europe as a whole. The range is some 270 miles (430 kilometres) long; it is barely six miles wide at its eastern end, but at its centre it reaches some 80 miles in width. At its western end it blends imperceptibly into the Cantabrian Mountains along the northern coast of the Iberian Peninsula. Except in a few places, where Spanish territory juts northward or French southward, the crest of the chain marks the boundary between the two countries, though the tiny, autonomous principality of Andorra lies among its peaks. The highest point is Aneto Peak, at 11,169 feet (3,404 metres), in the Maladeta ("Accursed") massif of the Central Pyrenees.

The Pyrenees long have been a formidable land barrier between Spain and Portugal on the Iberian Peninsula and the rest of Europe; as a consequence, these two countries traditionally have developed stronger associations with Africa than with the rest of Europe, and they have become tied to the sea. From Carlit Peak (9,584 feet) near the eastern limit of the Pyrenees to the peaks of Orhy and Anie, a succession of mountains rise nearly 9,800 feet; at only a few places, all well to the west, can the chain be crossed through passes lower than 6,500 feet. In both the lower eastern and northwestern sectors, rivers dissect the landscape into numerous small basins. The range is flanked on both sides by broad depressions—the Aquitaine and Languedoc to the north and the Ebro to the south—both receiving waters from the major rivers flowing out of the mountains, the Garonne of France and the major tributaries of the Ebro of Spain.

Cultural barrier

Physical features. *Geology.* The Pyrenees represent the geologic renewal of an old mountain chain rather than a more recent and vigorous mountain-building process that characterizes the Alps. Some 500 million years ago the region now occupied by the Pyrenees was covered with the folded mountains created during the Paleozoic era, called the Hercynian, of which the Massif Central in France and the Meseta Central in Spain are but two remnants. Although these other massifs have had a comparatively quiet history of internal deformation, or tectonism, since their emergence, the Pyrenean block was submerged in a relatively unstable area of the Earth's crust that became active about 225 million years ago.

The earliest formations, which were sediments severely folded over a granitic base, were submerged and covered by secondary sediments. They later were lifted once again into two parallel chains running to the north and south of the original Hercynian massif. These became the two



The Pyrenees mountain range.

zones of pre-Pyrenean ridges—of which the Spanish is the more fully developed—that are now great spurs of the main chain of the Pyrenees.

Under the forces of folding, the more recent and comparatively more plastic layers folded without breaking, but the original rigid base fractured and became dislocated. In the vicinity of the breaks, hot springs appeared and some metal-containing deposits formed. This upheaval affected chiefly the central and eastern regions. During this era, erosion continued incessantly, and, in the most exposed of the raised areas, weathering wore away the softer terrain and uncovered the old Hercynian sedimentary formations, occasionally reaching the deeper granitic bedrock.

Even today the old rocks, slates, schists, limestones transformed into marble (all of which come from old sediments transformed by great pressures and enormous heat), and granites of various kinds make up the spine, or axial zone, of the chain. The geologic phases of this zone, which rises and widens from west to east and ends by sinking, with a steep drop of nearly 9,800 feet, into the depths of the Mediterranean, have determined the evolution of the massif as a whole.

Physiography. The structure of the Pyrenees is characterized by patterns of relief and of underlying structure running in a north–south sequence (like the base rock); these alternate with depressions, some of which are the result of internal deformations, others of erosion of less resistant overlying deposits. In a cross section directly through the central area, where the tectonic activity reached its fullest width and development, it is possible to distinguish, from north to south, two strips of the comparatively recent pre-Pyrenean fold, one Spanish and one French, in juxtaposition with the axial massifs. An outer strip to the north consists of folds constituting the Petites Pyrénées. Cut into channels, they permit the passage of rivers. Nearer the middle of the range rise the Inner Ridges, represented by the mighty cliffs of the Ariège, which contain the primary, or granitic, axial zones. On the Spanish side the series is repeated in the opposite direction, but it is more highly developed and thicker. Thus the Interior Ridges—*e.g.*, Mount Perdido and the massif of Collarada—are sometimes higher than the neighbouring primary axial peaks. They are followed, to the south, by a broad, pre-Pyrenean, middle depression, with a succession of marine and continental deposits of varying hardness that constitute the valleys of such tributaries of the Ebro as the Aragón. This depression continues across the rest of the pre-Pyrenean ridges, among which are new secondary outcrops that form the fringe of Exterior Ridges and the northern rim of the depression of the Ebro; they are not, however, as thick or as important as the Interior Ridges.

North–south sequences of topography

From the structure of their relief and from the climatic conditions (especially on the south) that derive from the geographic situation of the chain, the Pyrenees have been divided into three natural regions: the Eastern (or Mediterranean), Pyrenees, the Central Pyrenees, and the Western Pyrenees. The different vegetation, the linguistic divisions of the people, and—to a point—certain ethnic and cultural distinctions appear to confirm this classification.

Drainage. The hydrographic system consists basically of series of parallel valleys that descend from the high peaks and from the passes. They are bordered by high, dividing ridges in a north–south direction, perpendicular to the axis of the chain. This type of valley produces short, torrential rivers that drop precipitously over short stretches; only seldom do these rivers flow, like the Aragón, through valleys that, as in the Alps, have both gentle slope and greater length. Their flow, extremely variable, especially on the southern side, is heavily influenced by the climate, as well as by the relief. Different maximum low waters occur in summer and winter; the spring, with maximum rain and melting snow, usually sees the greatest flows. In the Western Pyrenees and the northern zone, the rainfall pattern helps produce greater regularity; hence, flow is only slightly lower in summer. On the south a few torrential rivers are fed principally by melting snows, a few largely by rain, but most from a combination of sources.

The river patterns and flow have been important since antiquity in human use of both the land and the rivers—from the floating of timber rafts downstream, which can be done only in the spring, to harnessing waterpower for industry and irrigation on the southern side by means of dams. The torrential flow of many of the rivers is the cause both of the purity of the Pyrenean waters and of their excellence and richness as fishing streams.

The present Pyrenean glaciers, perhaps more frequent on the northern than on the southern slopes, have been reduced to high basins—cirques or hanging valleys—at elevations over 9,800 feet. During and after the great Ice Ages (*i.e.*, within the past 2.5 million years), however, especially in the Central and much of the Eastern Pyrenees, glaciers left widespread erosion and various important sediments. The present-day lower lakes and idyllic meadows with their winding rivulets are among their marks. Glacier tongues were also the main causes of the deep valleys containing the river system.

The fractured areas have many hot springs, both sulfurous and saline. The former are found throughout the axial massif, while the latter occur at the edges. These springs were popular in Roman times and reorganized and modernized toward the end of the 19th century. There are more than 20 famous spas on the French side; those in Spain are as numerous but are less fully exploited.

Climate. Major factors in the climate are the two abutting bodies of water and the extensive continental areas to the north and south. The Atlantic influence penetrates southward across the low peaks of the Western Pyrenees, as far south as Pamplona, Spain, tempering somewhat the differences of climate between the northern and southern slopes. This is not the case in the rest of the chain, especially the Central Pyrenees. The contrast in humidity between the French and Spanish sides is remarkable. To the north the oceanic influence, meeting no obstacles on the French plains of the Aquitaine, penetrates eastward and goes a little beyond the north–south watershed of the French rivers flowing into the Mediterranean. To the east the levanters, winds from the east and southeast, carry damp air from the Mediterranean, some of which falls as precipitation over the southeastern part of the eastern spurs. As a result, these regions are humid, while to the northeast the French depression of the Roussillon acquires Mediterranean characteristics.

Marine and terrestrial influences on climate

South of the Central Pyrenees the valley of the Ebro—which runs in a general northwest–southeast direction and is blocked by the southwest–northeast-trending Catalonian ranges near the eastern coast of Spain—acts as a "little continent." Hence, its climate is one of great thermal contrasts that are exaggerated by the generally high altitude of the Iberian Peninsula, but it is Mediterranean and unlike anything known in other European countries. Thus, the variegated climatic pattern of the Pyrenees ranges from the limpid, sunny atmosphere of the continental zone to the mild mists of the northwest and includes all transition stages in between.

Plant and animal life. Forms of life in the Pyrenees have some remarkable characteristics that cannot be explained merely by the influences of climate and soil. The historical vicissitudes of the chain and its isolation at the southwestern limit of the main European peninsula, far from the centres of dispersion and variation of the various species (including humans), have influenced the structure and character of its population.

In the northwest–southeast direction, the vegetation shows a marked and gradually decreasing oceanic influence; the contrary is the case with the Mediterranean influence from southeast to northwest. The exposure of the mountain surfaces and the conditions of local climate caused by mountain relief create special localized enclaves of all kinds. The most characteristic feature of the oceanic influence is the predominance of broad-leaved deciduous trees in the forests of the lower levels and the medium-height mountains, while the Mediterranean influence, represented by evergreen broad-leaved trees, not only is dominant in hot surroundings but also bears drought conditions better.

The variety of altitudinal vegetation shows itself in levels. From the medium-height mountain upward, the broad-leaved woods at about 5,200 feet are replaced by nee-

Vegetation zones

dled conifers that require less water. The subalpine level, sometimes as low as 6,500 feet but usually above 7,800 feet, gives way to the more sparsely covered pastures of the alpine level. This altitudinal scheme pertains in the vegetation east of the Orhy and Anie peaks. The oceanic influence, however, with its greater rainfall gives the west of the chain a different pattern. Broad-leaved deciduous beeches may be found as high as 5,850 feet, with some mix of the subalpine conifers, and there the high pastures are more resistant to damp and permanent snow. Overall the landscape is more like that of the high mountains of western Europe.

The Mediterranean influence expands through the entire valley of the Ebro, but it acquires marked signs of a more variable continental climate in the Central Pyrenees. There, great quantities of mountain pines, which are more drought-resistant, take the place of deciduous trees in the higher, colder, and drier parts of the medium and higher levels of the southern slopes.

Some groups among the fauna, such as the cave-dwelling animals and frogs and toads, represent a migratory wave that came from ancient Tyrrhenia—associated with Corsica and Sardinia—and displaced certain native European species, relegating them to the Cantabrian Mountains. The Pyrenean fauna is rich today, in larger herbivores as well as in the variety and abundance of predators. Some species, such as the wolf, lynx, and brown bear, have disappeared or had their numbers severely reduced in the northern Pyrenees, although the marmot has been successfully reintroduced. The southern Pyrenees, however, represent one of the last important reserves for wild European fauna driven out of sectors more heavily populated by humans. The present distribution and differentiation of large, warm-blooded animals is undoubtedly connected with the climate and the landscape, but the central-European origin of Pyrenean fauna is clear; for example, of the two species of desman (a semiaquatic member of the mole family), one inhabits the Pyrenees and the other the south central Soviet Union.

Similar comments may be made as to the origin of all cold-blooded animals as well as of the vegetation. Basic differentiations exist among the latter. Pyrenean flora of tropical origin differentiated without any ancient European competition as the new chain replaced the old Hercynian; flora of Arctic origin, brought southward during relatively recent ice ages, are represented by two different branches of orophiles, or plants adapted to mountain life, from central Europe and from Siberia. Other orophiles have long been differentiated, but they are of Mediterranean origin and are dominant in the drier, sunnier parts of the southern slopes. An Atlantic group of flora predominates in the Western Pyrenees.

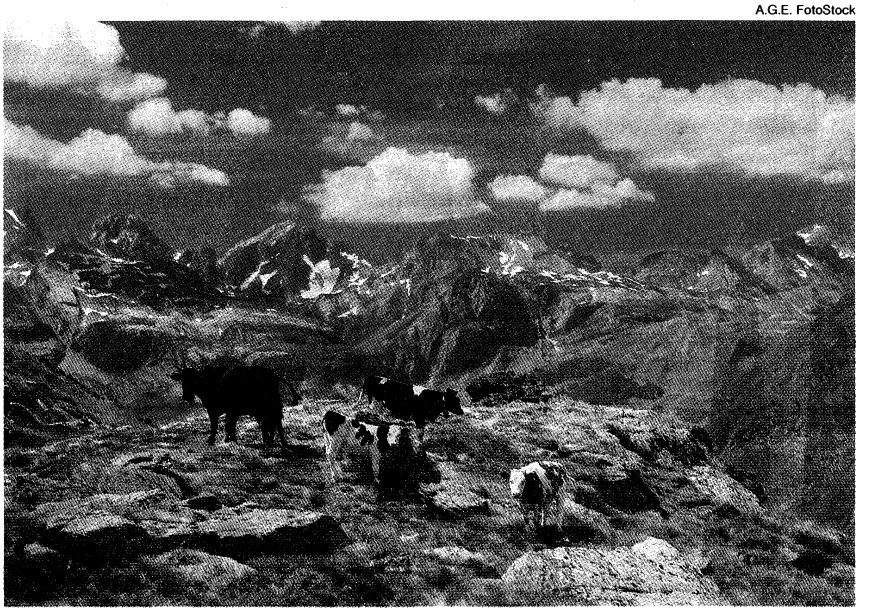
The people and economy. The Pyrenees are the home of a variety of peoples, including the Andorrans, Catalans, Béarnais, and Basques. Each speaks its own dialect or language, and each desires to maintain and even augment its own autonomy while at the same time acknowledging a general unity among Pyrenean peoples. Of these groups, only the Andorrans have anything approaching a sovereign state, and even then Andorra is an autonomous principality with close ties with both Spain and France. The Basques, perhaps the best-known Pyrenean people, speak a language that is non-Indo-European and have a long tradition of fiercely defending their autonomy. The Basques

The people of the Pyrenees traditionally have depended on agriculture and livestock raising for their livelihood. The factors that influenced the development of Pyrenean flora also influenced traditional land usage, the kind of crops raised, and the farming system of each district. Typical Mediterranean products such as wines, vegetables, and fruits predominate in the Eastern Pyrenees and at the foot of the chain's southern slope, while in the Western and Central Pyrenees, with their abundant rainfall, potatoes, sweet corn, and forage crops are grown. Livestock breeding, the other essential element of the traditional economy, consists of a seasonal process of moving flocks of sheep and cows up and down the mountains and also using as well as possible the meadows of the valley bottoms and the pastures of the higher altitudes, depending on the snow cover. Frequently in winter, the livestock herds travel far from the Pyrenees, moving to the plains of the Ebro, near the Mediterranean Sea in Languedoc, or on the moors of Aquitaine.

This traditional organization—in which the common exploitation of forest areas for timber also played a large part—has been disappearing slowly. Most farmers are now elderly, and few young people have been willing to settle into the old ways. Gradually, the less fertile plots have been deserted, and the landscape has become dotted by patches of brooms and brackens and plantings of resinous trees. Even local breeds of sheep or cows have been superseded by imported breeds, which perhaps are more profitable but are less adapted to the climate and the relief. Except for such areas as the Basque Country of Spain and the Roussillon region of France, the agriculture of the Pyrenees is in serious decline.

The growing weakness of the Pyrenean agriculture has not been matched by growth in industry. Although the Pyrenees offer considerable hydroelectric potential, the mining of some resources, and an appreciable and diverse supply of wood, most of the mills (steel and paper) and factories (textiles, chemicals, and shoes) established in the 19th and 20th centuries have faced the threat of closing. Except in the two extremities of the chain, most of the

A.G.E. FotoStock



Cows grazing high in the Central Pyrenees, Huesca province, Spain.

industries are far from any major transportation routes. Scarcely any railroads and no major highways traverse the region, although an express highway is slowly being built between Toulouse, Fr., and Barcelona, Spain. Financed by foreign capital and dependent on the aid of the Spanish and French governments, the remaining factories face an uncertain future.

Perhaps the policies that have been formulated since 1980 by the two Pyrenean countries to develop and protect the mountains may slow down the exile of Pyrenean peoples, who have seen their massif transformed by the tremendous increase in tourism. Although a boon to the local economy, the crowds of people seeking winter sports, summer sojourns, hunting and fishing, and visits to the national parks of the Central Pyrenees have also contributed to the abandonment of traditional ways of life.

Tourism

Study and exploration. For centuries a general lack of knowledge about the Pyrenees permitted repetition of the errors and misconceptions about the mountains that had been propounded by such authors of antiquity as Diodorus Siculus of Sicily and the Greek geographer-historian Strabo (both 1st century BC). In 1582 the first explorations were made, followed by botanical works from the academies at Montpellier-de-Médillan, Fr., and by other studies, including those of the 18th-century Swiss pioneer alpinist Horace Bénédict de Saussure. The earliest military map of the region dates from 1719, while early topographical studies were the bases of frontier treaties.

In the 19th century the first topographical and geologic maps were made of the mountains, the latter beginning a series of geologic interpretations and controversies among French and Spanish scientists. German studies added to the interpretive geology, but only in 1933 was the first study made that was based on modern research methods. Since World War II, scientists and scholars from universities, technical institutes, and national councils for research in France and Spain have thoroughly explored the Pyrenees and have produced a wealth of knowledge about the massif. (F.O./Ed.)

URAL MOUNTAINS

The major part of the traditional boundary between Europe and Asia, the Ural Mountains are a rugged spine across the middle of the Soviet Union, running more than 1,300 miles (2,100 kilometres) from the fringe of the Arctic in the north to the bend of the Ural River in the south. The low, severely eroded Pay-Khoy Ridge forms a fingerlike extension to the northern tip of the Urals proper, with the long curve of Novaya Zemlya forming an insular extension separating the Barents and Kara seas. The Mugodzhar Hills, a broad, arrowhead-shaped southerly extension, form the divide between the Caspian and Aral basins. The north–south course of the Urals is relatively narrow, varying from about 20 to 90 miles in width, but it cuts across the vast latitudinal landscape regions of the Eurasian landmass, from Arctic waste to semidesert; the Urals also are part of the Ural Economic Region, a highly developed industrial complex closely tied to the mineral-rich Siberian region, and are the home of peoples with roots reaching deep into history.

Physical features. *Physiography.* The Urals divide into five sections. The northernmost Polar Urals extend some 240 miles from Mount Konstantinov Kamen in the northeast to the Khulga River in the southeast; most mountains rise to 3,300–3,600 feet (1,000–1,100 metres) above sea level, although the highest peak, Mount Payer, reaches 4,829 feet. The next stretch, the Nether-Polar Urals, extends for more than 140 miles south to the Shchugor River. This section contains the highest peaks of the entire range, including Mount Narodnaya (6,217 feet [1,895 metres]) and Mount Karpinsk (6,161 feet). These first two sections are typically Alpine and are strewn with glaciers and heavily marked by permafrost. Farther south come the Northern Urals, which stretch for more than 340 miles to the Usa River in the south; most mountains top 3,300 feet, and the highest peak, Mount Telpos-Iz, rises to 5,305 feet. Many of the summits are flattened, the remnants of ancient peneplains (eroded surfaces of large area and slight relief) uplifted by geologically recent tectonic movements. In the north, intensive weathering has resulted in vast "seas of stone" on mountain slopes and summits. The lower Central Urals, extending more than 200 miles to the Ufa River, rarely exceed 1,600 feet, though the highest peak, Mount Sredny Baseg, rises to 3,261 feet. The summits are smooth, with isolated residual outcrops. The last portion, the Southern Urals, extends some 340 miles to the westward bend of the Ural River and consists of several parallel ridges rising to 3,900 feet and culminating in Mount Yamantau, 5,380 feet; the section terminates in the wide uplands (less than 2,000 feet) of the Mugodzhar Hills.

The Urals' five sections

Tass/Sovfoto

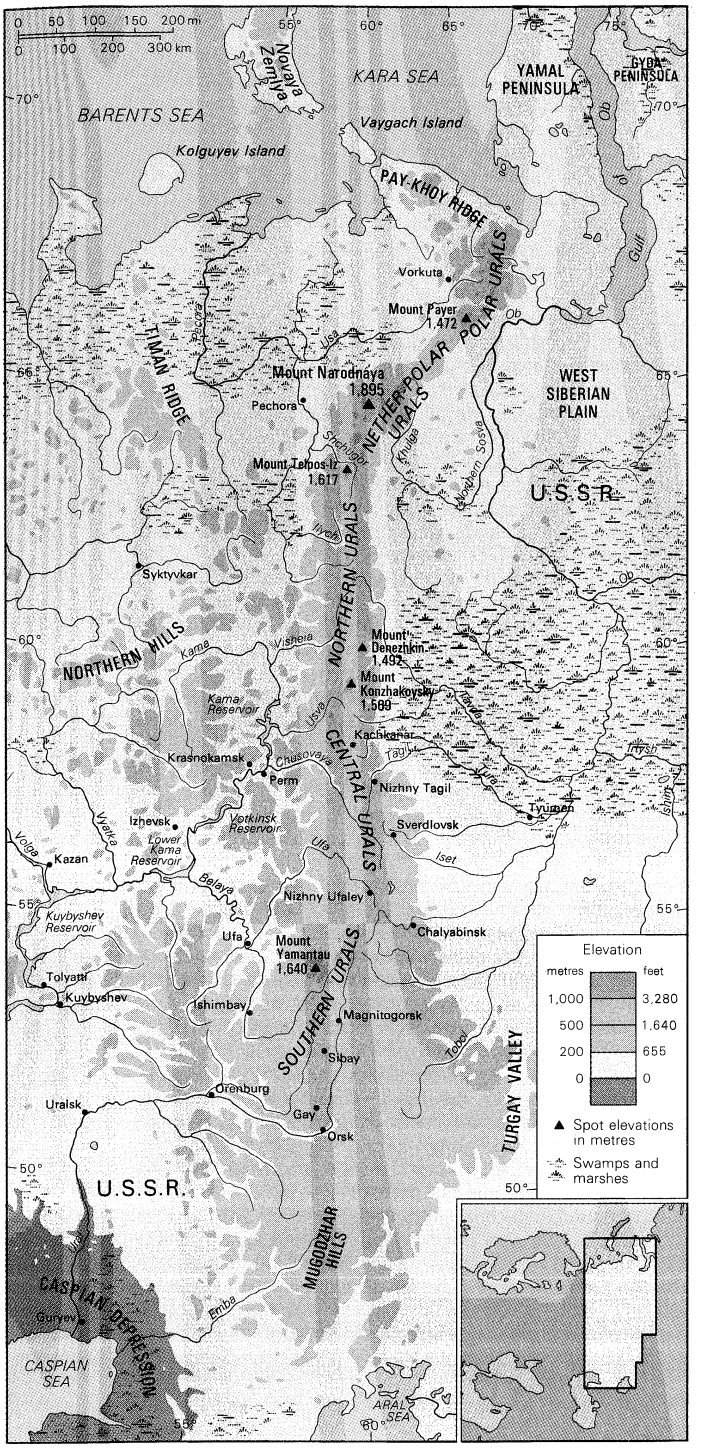
Nurgush Range, Southern Ural Mountains, Russian S.F.S.R.

The rock composition helps shape the topography: the high ranges and low, broad-topped ridges consist of quartzites, schists, and gabbro, all weather-resistant. Buttes are frequent, and there are north–south troughs of limestone, nearly all containing river valleys. Karst topography is highly developed on the western slopes of the Urals, with many caves, basins, and underground streams. The eastern slopes, on the other hand, have fewer karst formations; instead, rocky outliers rise above the flattened surfaces. Broad foothills, reduced to peneplain, adjoin the Central and Southern Urals on the east.

Geology. The Urals date from the structural upheavals of the Hercynian orogeny (about 250 million years ago). About 280 million years ago there arose a high mountainous region, which was eroded to a peneplain. Alpine folding resulted in new mountains, the most marked upheaval being that of the Nether-Polar Urals. In the watershed region lies the Ural-Tau Anticlinorium (a rock formation of arches and troughs, itself forming an arch), the largest in the Urals, and in the Southern Urals, west of it, is the Bashkir Anticlinorium. Both are composed of layers (sometimes four miles thick) of ancient metamorphic (heat-altered) rocks—gneisses (metamorphic rocks separable into thin plates), quartzites, and schists—that are between 570 and 395 million years old.

The western slope of the Urals is composed of middle Paleozoic sedimentary rocks (sandstones and limestones) that are about 350 million years old. In many places it descends in terraces to the Cis-Ural depression (west of the Urals), to which much of the eroded matter was carried during the late Paleozoic (about 300 million years ago). Here there are widespread karst (a starkly eroded limestone region) and gypsum, with large caverns and subterranean streams. On the eastern slope, volcanic layers alternate with sedimentary strata, all dating from middle Paleozoic times. These rocks compose the Tagil-Magnitogorsk Synclinorium (a group of rock arches and troughs, itself forming a trough), the largest in the Urals. In the Central and Southern Urals the eastern slope blends into

Sedimentary rocks of the western slope



The Ural Mountains.

broad peneplained foothills, where there are frequent outcrops of granite and often fantastically shaped buttes. To the north the peneplain is buried under the loose, easily pulverized deposits of the West Siberian Plain.

Drainage. The rivers flowing down from the Urals drain into either the Arctic Ocean or the Caspian Sea. The Pechora River, which drains the western slope of the Polar, Nether-Polar, and part of the Northern Urals, empties into the Barents Sea. Its largest tributaries are the Ilych, Shchugor, and Usa. Almost all the rivers of the eastern slope belong to the Ob River system, emptying into the Kara Sea. The largest are the Tobol, the Iset, the Tura, the Tavda, the Severnaya Sosva (Northern Sosva), and the Lyapin. The Kama (a tributary of the Volga) and the Ural rivers belong to the drainage basin of the Caspian Sea. The Kama collects water from a large area of the western slope: the Vishera, Chusovaya, and Belaya all empty into it. The Ural River, with its tributary the Sakmara, flows along the Southern Urals.

The location and character of the Urals' rivers and lakes are closely connected with the topography and climate. In their upper reaches many rivers flow slowly through the mountains in wide, longitudinal troughs. Later they change to a latitudinal direction, cut through the ridges in narrow valleys, and descend to the plains, particularly in the Northern and Southern Urals. The main watershed does not correspond with the highest ridges everywhere. The Chusovaya and Ufa rivers of the Central and Southern Urals, which later join the Volga drainage basin, have their sources on the eastern slope.

The rivers on the western slope carry more water than those of the east, particularly in the Northern and Nether-Polar Urals; the slowest rate of flow is on the eastern slope of the Southern Urals, reflecting intense evaporation as well as low precipitation. In winter the rivers freeze for five months in the south and for seven months in the north.

There are many lakes, especially on the eastern slope of the Southern and Central Urals. The largest are Uvildy, Itkul, Turgoyak, and Tavatuy. On the western slope are many small karst lakes. In the Polar Urals, lakes occur in glacial valleys, the deepest of them being Lake Bolshoye Shchuchye, at 446 feet deep. Medicinal muds are common in a number of the lakes, such as Moltayevo, and spas and sanatoriums have been established.

Climate. The climate is of the continental type, marked by temperature extremes that become increasingly evident both from north to south and from west to east. The Pay-Khoy Range and the Polar Urals enjoy the moderating influence of the Arctic and the North Atlantic oceans, particularly in winter. In the Mugodzhar Hills and the Southern Urals there are summer winds of hot, dry air from Central Asia. Winds are for the most part westerly and bring precipitation from the Atlantic Ocean. In spite of their low elevation, the mountains exert a considerable influence on the moisture distribution, and the western slope receives more moisture than the eastern. Precipitation is particularly heavy on the western slope of the Nether-Polar and Northern Urals, as high as 40 inches (1,000 millimetres). Northward and southward precipitation diminishes to about 18 inches. On the eastern slope there is less moisture (about 12 inches) and snow. Annual snow depth on the western slope averages 35 inches and on the eastern, 18 inches. Maximum precipitation occurs in the summer, for the cold, dry air of the Siberian anticyclone is powerful in winter. The eastern slope is particularly chilled, and winter lasts longer than summer throughout the Urals. In January the average temperature in the north is −6° F (−21° C), and in the south the average is 5° F (−15° C). Average temperatures in July vary more, between 50° F (10° C) in the north and 72° F (22° C) in the south.

Precipitation

Plant life. The Urals pass through several vegetation zones, with the northern tundra giving way to vast mixed forests, while still farther south is the steppe, culminating in semidesert around the Mugodzhar Hills. Feather grass and meadows predominate on the chernozems (black earth) and dark chestnut soils (a characteristic steppe soil). Other characteristic growths are clover, fescue (a pasture grass), and timothy (a grass grown for hay). South of the Ural River the steppes give way to wormwood and semidesert growths on light chestnut soils (again typical steppe soil), which are highly saline in places.

The forest landscapes of the Urals are varied. The more humid western foothills of the Southern Urals are covered mostly by mixed forests growing on a gray mountain-forest type of soil. There, such broad-leaved species as oak, small-leaf linden, and elm are mixed with Siberian fir and Siberian spruce. The broad-leaved forests extend to 2,100 feet, above which conifers appear. On the eastern slope there are no broad-leaved trees except the linden, and magnificent pine forests with some larches are widespread.

Farther to the north, in the Central Urals, taiga forests of spruce, fir, pine, and larch grow on the mountain, podsolic soils. In the more northerly regions, dark coniferous species are common, and, in the Northern Urals,

the Siberian cedar is widespread. There forests climb to 2,800 feet or so, above which is a narrow belt of larch and birch, trailing off to mountain tundra. In the Nether-Polar and Polar Urals the forest yields to mountain tundra at elevations as low as 1,300 feet. Whereas moss tundra is generally found on the more humid western slope, lichen tundra is common on the eastern. There are numerous sphagnum moss marshes on both slopes. Only brushwood and moss–lichen tundra grows on the Pay-Khoy Ridge.

Animal life. There are no specifically mountain animals in the Urals, primarily because of the low elevations and easy accessibility, and fauna differs little from that of the adjacent areas of eastern Europe and western Siberia. The most valuable animal of the tundra is the Arctic fox. Ob lemmings, snowy owls, tundra partridge, and reindeer are other inhabitants, though the latter are few. Many wild ducks, geese, and swans breed there in summer. But the richest and most varied fauna in the Urals, such as the brown bear, lynx, wolverine, and elk, are found in the forested zones. Some have valuable furs: the sable (in the Northern Urals), ermine, fox, marten (in the Southern Urals), Siberian weasel, and squirrel. In the taiga forests there are such birds as the wood grouse, black grouse, capercaillie (another member of the grouse family), cuckoo, and hazel hen (a woodland grouse). In the mixed, broad-leaved forests of the Southern Urals' western slopes live roe deer, badgers, and polecats, as well as many birds typical of the European part of the Soviet Union, such as the nightingale and oriole. The commonest animals of the steppes and semideserts are rodents, including susliks (a type of ground squirrel), jerboas (a social, nocturnal, jumping rodent), and other agricultural pests. Reptiles include the common adder and grass snake. The rivers and lakes of the Northern Urals abound in fish, the most valuable being the nelma (related to the whitefish), common salmon, grayling, and sea trout. Farther south, in the densely populated and industrial regions, animal life is less abundant.

Similarity of fauna in adjacent regions

The vigorous economic development and growth in population that have occurred in the Urals in the 20th century have altered considerably the chain's landscape and the abundance of wildlife. Conservation measures during the Soviet period have included establishing national nature preserves such as Pechoro-Ilych in the Northern Urals, Basegi and Visim in the Central Urals, and Ilmen and Bashkir in the Southern Urals.

The people. Human habitation of the Urals dates to the distant past. The Nenets (Nentsy) are a Samoyed people of the Pay-Khoy region, and their language belongs to the Samoyedic group of languages, which is widespread throughout northern Siberia. Farther south live the Komi, Mansi (Voguls), and Khants (Ostyaks), who speak a tongue belonging to the Ugric group of the Finno-Ugric languages. The most numerous indigenous group, the Bashkirs, long settled in the Southern Urals, speak a tongue related to the Turkic group. Some Kazakhs live in the Mugodzhar Hills. Most of these formerly nomadic peoples are now settled. The Nenets, Komi, Mansi, and Khants are virtually the only inhabitants in the highest parts of the Urals, especially in the north, where they have preserved their traditional ways of life, raising reindeer, hunting, and fishing. The Bashkirs are excellent horse breeders. The indigenous peoples, however, now constitute only about one-fifth of the total population of the Urals; the great majority are Russians. The Russian population is concentrated primarily in the Central and Southern Urals, and most people live in cities (notably Sverdlovsk, Chelyabinsk, Perm, and Ufa) and work in industries. Agricultural populations predominate in the steppe region of the Southern Urals, where conditions are favourable for wheat, potatoes, and other crops.

The economy. The Urals are extremely rich in mineral resources, with variations on the eastern and western slopes according to geologic structure. Ore deposits, for example, notably magnetite, predominate on the eastern slope, where contact (the surface where two different rock types join) deposits are found, as at Vysokogorsk and Mount Blagodat, as well as magmatic deposits (formed from liquid rock), as at Kachkanar. Some of the ore deposits, such as the magnetite ores at Magnitogorsk, are exhausted or nearly depleted. Sedimentary deposits are of less importance. Some ores contain alloying metals—vanadium, a gray-white, resistant element, and titanium—as impurities. The largest copper ore deposits are at Gay and Sibay, and nickel ores are found at Ufaley. There are also large deposits of bauxite, chromite, gold, and platinum.

Mineral resources

Among the nonmetallic mineral resources of the eastern slope are asbestos, talc, fireclay, and abrasives. Gems and semiprecious stones have long been known: they include amethyst, topaz, and emerald. Among the western deposits are beds of potassium salts on the upper Kama River and petroleum and natural gas deposits in the Ishimbay and Krasnokamsk areas. Bituminous coal and lignite are mined on both slopes. The largest deposit is the Pechora bituminous coalfield in the north.

The vast forests of the Urals are also of great economic importance: not only do they yield valuable wood, but they also regulate the flow of the rivers and shelter many of the valuable fur animals. Agriculture is significant mainly in the eastern steppe region of the Southern Urals. Much of the land there has been plowed up and converted to arable lands, and in large areas wheat, buckwheat, millet, potatoes, and vegetables are grown.

Because of its wealth of mineral resources, the leading industries in the Urals are mining, metallurgy, machine building, and chemicals. Of national importance are the metallurgical plants at Magnitogorsk, Chelyabinsk, and Nizhny Tagil; chemical plants at Perm, Ufa, and Orenburg; and large-scale engineering at Sverdlovsk.

Study and exploration. The existence of the Riphean and Hyperborean mountains at the eastern fringe of Europe in antiquity was regarded as being more mythical than real. Not until the 10th century AD does the first mention of the Urals occur, in Arabic sources. At the end of the 11th century the Russians discovered the northernmost part of the Urals, but they did not complete the discovery of the entire range until the beginning of the 17th century, when the mineral wealth of the Urals was discovered. The first geographic survey of the chain was made in the early 18th century by the Russian historian and geographer Vasily N. Tatishchev, who undertook the survey for Peter I the Great. Systematic extraction of iron and copper ore also began at that time, and the Urals rapidly became one of the largest industrial regions of Russia.

Scientific studies of the Urals

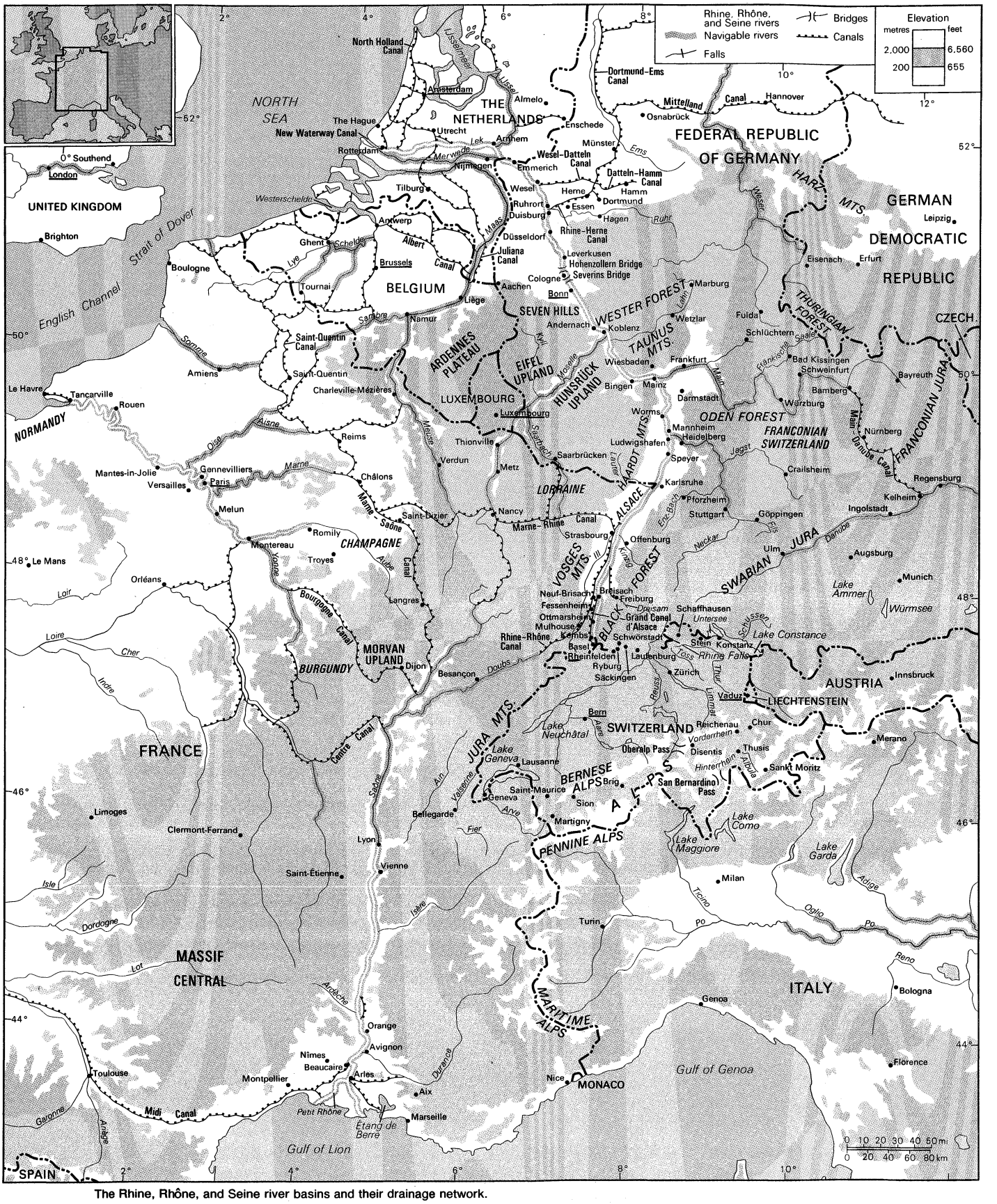
The first serious scientific study of the Urals was made in 1770–71. Scholars studying the Urals during the 19th century included several Russian scientists, such as the geologist A.D. Karpinsky, the botanist P.N. Krylov, and the zoologist L.P. Sabaneev, and also such prominent foreign scholars as the German naturalist Alexander von Humboldt and the English geologist Sir Roderick Murchison, who compiled the first geologic map of the Urals in 1841. Much work has been done in the Soviet period on geologic structure and associated mineral resources. (Y.V.Y.)

Western European drainage systems

RHINE RIVER

Culturally and historically one of the great rivers of Europe and among the most important arteries of industrial transport in the world, the Rhine River flows 865 miles (1,390 kilometres) from east central Switzerland north and west to the North Sea, into which it drains through The Netherlands. The German spelling is Rhein; Dutch, Rijn; French, Rhin—all derived from the Latin, Rhenus. An international waterway since the Treaty of Vienna in 1815, it is navigable overall for some 540 miles, as far as Rheinfelden on the Swiss-West German border. Its catchment area, including the delta area, exceeds 85,000 square miles (220,000 square kilometres).

The Rhine has been a classic example of the alternating roles of great rivers as arteries of political and cultural unification and as political and cultural boundary lines. The river has also been enshrined in the literature of its lands, especially of Germany, as in the famous epic *Nibelungenlied.* Since the time when the Rhine valley became incorporated into the Roman Empire, the river has been one of Europe's leading transport routes. Until the



The Rhine, Rhône, and Seine river basins and their drainage network.

19th century the goods transported were of high value but relatively small in volume, but since the second half of the 19th century the volume of goods conveyed on the river has increased greatly. The fact that cheap water transport on the Rhine helped to keep prices of raw materials down was the main reason the river became a major axis of industrial production: one-fifth of the world's chemical industries are now manufactured along the Rhine. The river was long a source of political dissension in Europe, but this has given way to international concern for ecological safeguards as pollution levels have risen; some 6,000 toxic substances have been identified in Rhine waters.

No other river in the world has so many old and famous cities on its banks—Basel, Switz.; Strasbourg, Fr.; and Worms, Mainz, and Cologne, W.Ger., to name a few—but there are also such industrial cities as Ludwigshafen and Leverkusen in West Germany, which pollute the waters and mar the scenic attraction of the riverbanks. Nonetheless, the middle Rhine (the section between the West German cities of Bingen and Bonn), with such steep rock precipices as the Lorelei crag and numerous castles, still presents breathtaking vistas and attracts tourists. This is the Rhine of legend and myth, where the medieval Mouse Tower (Mausturm) lies at water level near Bingen, and the castle of Kaub stands on an island in the river. The Alpine section of the Rhine lies in Switzerland, and below Basel the river forms the boundary between West Germany and France, as far downstream as the Lauter River. It then flows through West German territory as far as Emmerich, below which its many-branched delta section epitomizes the landscapes characteristic of The Netherlands.

Source

Physical features. *Physiography.* The Rhine rises in two headstreams high in the Swiss Alps. The Vorderrhein emerges from Lake Toma at 7,690 feet, near the Oberalp Pass in the Central Alps, and then flows eastward past Disentis to be joined by the Hinterrhein from the south at Reichenau above Chur. (The Hinterrhein rises about five miles west of San Bernardino Pass, near the Swiss-Italian border, and is joined by the Albula River below Thusis.) Below Chur, the Rhine leaves the Alps to form the boundary first between Switzerland and the principality of Liechtenstein and then between Switzerland and Austria, before forming a delta as the current slackens at the entrance to Lake Constance. In this flat-floored section the Rhine has been straightened and the banks reinforced to prevent flooding. The Rhine leaves the lake via its Untersee arm. From there to its bend at Basel, the river is called the Hochrhein ("High Rhine") and defines the Swiss-German frontier, except for the area below Stein am Rhein, where the frontier deviates so that the Rhine Falls at Schaffhausen are entirely within Switzerland. Downstream, the Rhine flows swiftly between the Alpine foreland and the Black Forest region, its course interrupted by rapids, where—as at Laufenburg, Säckingen, and Schwörstadt—barrages (dams) have been built. In this stretch the Rhine is joined by its Alpine tributaries, the Thur, Töss, Glatt, and Aare, and by the Wutach from the north. The Rhine has been navigable between Basel and Rheinfelden since 1934.

Below Basel the Rhine turns northward to flow across a broad, flat-floored valley, some 20 miles wide, held between, respectively, the ancient massifs of the Vosges–Black Forest and the Haardt–Oden Mountains. The main tributary from Alsace is the Ill, which joins the Rhine at Strasbourg, and various shorter rivers, such as the Dreisam and the Kinzig, drain from the Black Forest. Downstream, the regulated Neckar, after crossing the Oden Mountains in a spectacular gorge as far as Heidelberg, enters the Rhine at Mannheim; and the Main leaves the plain of lower Franconian Switzerland for the Rhine opposite Mainz. Until the straightening of the upper Rhine, which began in the early 19th century, the river described a series of great loops, or meanders, over its floodplain, and today their remnants, the old backwaters and cutoffs near Breisach and Karlsruhe, serve to mark the former course of the river.

The middle Rhine

The middle Rhine is the most spectacular and romantic reach of the river. In this 90-mile stretch the Rhine has cut a deep and winding gorge between the steep, slate-covered slopes of the Hunsrück Mountains to the west and the Taunus Mountains to the east. Vineyards mantle the slopes as far as Koblenz, where the Moselle River joins the Rhine at the site the Romans called Confluentes. On the right bank, the fortress of Ehrenbreitstein dominates the Rhine where the Lahn tributary enters. Downstream the hills recede, the foothills of the volcanic Eifel region lying to the west and those of the Wester Forest to the east. At Andernach, where the ancient Roman frontier left the Rhine, the basaltic Seven Hills rise steeply to the east of the river, where, as the English poet Lord Byron put it, "the castle crag of Dachenfels frowns o'er the wide and winding Rhine."

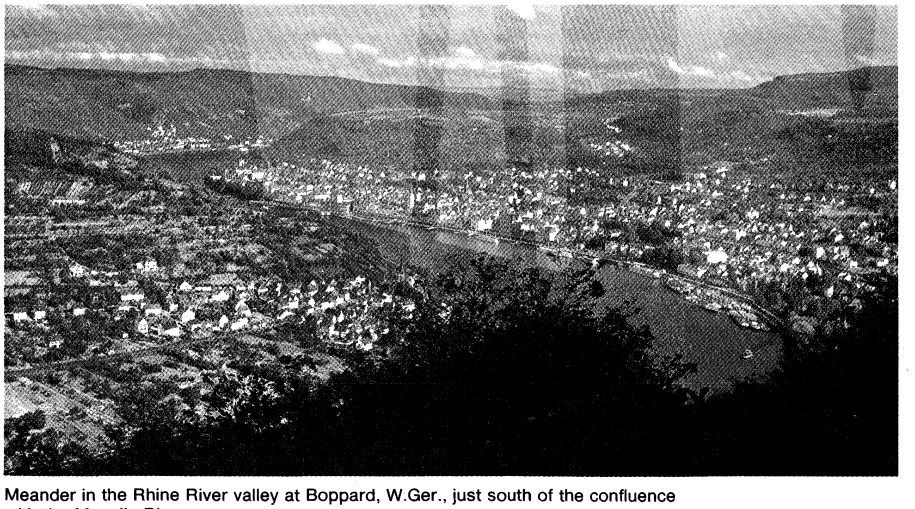
Below Bonn the valley opens out into a broad plain, where the old city of Cologne lies on the left bank of the Rhine. There the river is spanned by the modern Severin Bridge and by the rebuilt Hohenzollern railway bridge, which carries the line from Aachen to Düsseldorf and the Ruhr industrial region. Düsseldorf, on the right bank of the Rhine, is the dominant business centre of the North Rhine–Westphalia coalfield. Duisburg, which lies at the mouth of the Ruhr River, handles the bulk of the waterborne coal and coke from the Ruhr as well as imports of iron ore and oil.

The delta region

The last section of the Rhine lies below the frontier town of Emmerich in the delta region of The Netherlands. There the Rhine breaks up into a number of wide branches, such as the Lek and Waal, farther downstream called the Merwede. With the completion of the huge Delta Project in 1986—constructed to prevent flooding in the southwestern coastal area of The Netherlands—all main branches of the Rhine were closed off; sluices and lateral channels now allow river water to reach the sea. Since 1872, however, the New Waterway Canal, constructed to improve access from the North Sea to Rotterdam, has been the main navigation link between the Rhine and the sea; along this canal was built Europoort, one of the world's largest ports. (A.F.A.M./K.A.Si.)

Hydrology. The Alpine Rhine—with its steep gradient, high runoff coefficient (80 percent of the precipitation in its catchment area), pronounced winter minimum, high water in spring from snowmelt, and high early summer maximum resulting from heavy summer rains—has a

J. Allan Cash



Meander in the Rhine River valley at Boppard, W.Ger., just south of the confluence with the Moselle River.

characteristic Alpine regime. Although variations in flow are evened out by Lake Constance, which is fed by upland streams as well as by the Rhine (and which also acts as a filter), they are increased again by the confluence with the Aare, which on an average carries more water than the Rhine. Below Basel, however, the tributaries from the uplands, with their spring maximums at higher and winter maximums at lower elevations, increasingly moderate the unbalance. Thus, at Cologne the average deviations from mean flow are slight, and the regime is favourable to navigation. Winters in the navigable regions of the river, moreover, are generally mild, and the Rhine freezes only in exceptional winters.

The economy. As a commercial artery, the Rhine is unrivaled among the world's rivers, historically as well as in the amount of traffic carried. The Romans maintained a Rhine fleet, and the importance of the river increased enormously with the rise of medieval trade, which relied on water transport wherever possible because of the poor roads. The rock barrier of the gorge at Bingen divided navigation into two sections: predominantly upstream traffic by seagoing vessels to Cologne and predominantly downstream movement of commodities—brought first across the Alpine passes—from Basel to Mainz and Frankfurt am Main. After about 1500, navigation declined because of reorientation of trade toward the Atlantic and political disintegration of the Rhineland. The rise of modern navigation began in the 19th century, and its present magnitude is attributable largely to four factors: removal of political restrictions on navigation, physical improvements to the navigation channel, canalization of the Rhine's hinterland, and increasing industrialization of the riparian countries.

Navigation agreements

The principle of free navigation on the Rhine was agreed upon by the Congress of Vienna in 1815 and was put into effect by the Mainz Convention of 1831, which also established the Central Commission of the Rhine. This first treaty was simplified and revised in the Mannheim Convention of 1868, which, with the extension in 1918 of all privileges to ships of all countries and not merely the riverine states, remains (broadly speaking) in force.

Navigational improvements. Historically, two sections presented serious handicaps to navigation: the rock barrier at Bingen and the southern upper Rhine. At Bingen two navigation channels were blasted out in 1830–32; canalization of the upper Rhine by confining it within an artificial bed and straightening its course was undertaken in 1817–74. In neither case were the resulting improvements entirely satisfactory, but the channels at Bingen were doubled in width and deepened, thus eliminating the need for a pilot. Navigation on the upper Rhine, despite the further improvements made after 1907, suffers from seasonal variations of flow and the swift current.

To improve navigation and to procure hydroelectric power, France (by the Treaty of Versailles) obtained the right to divert Rhine water below Basel into a canal that was to rejoin the Rhine at Strasbourg. Construction of the first section of this Grand Canal d'Alsace, designed to take vessels of 1,500 tons, was completed with the building of a dam at Kembs in 1932 and greatly improved navigation. Construction was resumed after World War II, but in a treaty (1956) France, in return for West German agreement to the canalization of the Moselle, consented to terminate the canal at Neu Breisach. The remaining four of a total of eight dams utilize Rhine water by the construction of canal loops only.

Below Basel the Huningue branch of the Rhine–Rhône Canal leads to Mulhouse, where it meets the main arm of that waterway, which joins the Rhine at Strasbourg. The Rhine–Rhône Canal (1810–33) is navigable by 300-ton craft and carries only moderate traffic. More important, although no larger, is the Rhine–Marne Canal (1838–53), which also joins the Rhine at Strasbourg.

The Neckar is canalized through Stuttgart as far as Plochingen and the Main as far as Bamberg. There, the completed northern portion of the Main–Danube Canal leads south to Nürnberg, which has become an important port. A treaty signed in 1956 between West Germany, France, and Luxembourg provided for canalization of the Moselle from Koblenz to Thionville (170 miles), which was completed in 1964. The Lahn also is canalized for small (200-ton) craft for 42 miles.

In the Ruhr region the Ruhr itself (except for the last seven miles) and the Lippe are not used as waterways. Their place is taken by the Rhine–Herne Canal, completed in 1916 between Duisburg and Herne and linking the Rhine through the Dortmund–Ems Canal with the German North Sea coast and through the Mittelland Canal with the waterways of central and eastern Germany and eastern Europe; and by the less important Wesel–Datteln–Hamm Canal (1930), which runs parallel to the lower course of the Lippe. The Rhine–Herne Canal's capacity for craft of 1,350 tons became the standard both for the minimum capacity of canals built since World War II and for barges. Nearer the Rhine's mouth, the Merwede Canal (enlarged 1952) south of Amsterdam provides another route to the sea for ships displacing as much as 4,300 tons.

Growth of the Rhine as a waterway

Traffic. Three factors were important in the rise of traffic on the Rhine. First, the political impediments to free navigation—particularly the approximately 200 toll stations along the course of the river—were removed by the Congress of Vienna of 1815. Second, the means of transport were improved by the introduction of steam-powered, and later diesel-powered, tugs; prior to the mid-19th century, barges moving upstream were towed either by teams of horses or gangs of men. Third, the waterway itself was improved, the stages of which are discussed above.

The first steamship voyage on the Rhine was made from London to Koblenz in 1817, but this was a solitary event. The harbour installations of Mannheim were opened in 1840, and for almost a century this was the effective head of navigation. Although Basel had been reached by a steamship by 1832, its development as a Rhine port started a century later. Despite the improvement of the navigation and means of transport, there was at first little growth in the volume of transport. Increase came with the rise of modern industry in the 19th century, which necessitated the bulk movement of coal, ore, building materials, raw material for the chemical industry, and (since about 1950) oil. Although coal and ore transport declined, there was an overall increase in the volume of transport until the mid-1960s; since then, however, freight tonnage has decreased to about a third of its former level.

The mode of transport from 1840 onward was by tugs towing a number of barges. Development after 1945 involved initially the introduction of self-propelled barges and subsequently the introduction of push tugs, whereby one tug can propel four-barge units and thus save labour costs. An increase in the traffic volume was also effected by the introduction of radar navigation in the 1950s, which made round-the-clock operation possible. There is also regular passenger service on the Rhine during summer, especially the middle Rhine section and from Rotterdam to Basel, but this is almost exclusively for tourists.

History. The effects of rivers on the regions through which they flow tend to alternate between trends toward unifying the regions culturally and politically and making a political boundary of the river. Of this phenomenon the Rhine is a classic example. During prehistoric times the same culture groups existed on both banks; similarly, in early historic times Germanic tribes settled on either side of its lower and Celts alongside its upper course. Although bridged and crossed by Julius Caesar in 55 and 53 BC, the Rhine became for the first time, along its course from Lake Constance to its mouth at Lugdunum Batavorum (Leiden, Neth.), a political boundary—that of Roman Gaul. This division did not endure for long, because under the emperor Augustus the provinces of Germania Superior and Germania Inferior were established on the other side of the Rhine, and south of Bonna (Bonn) the boundary of the Roman Empire was marked by the limes (Roman fortified frontier) well east of the river. Nevertheless, because the Rhine had been the boundary of Gaul for a time, it resulted in later claims by France, esteeming itself the successor to Gaul, to the Rhine as its natural boundary. When the Western Roman Empire disintegrated, the Rhine was crossed along its entire length by Germanic tribes (AD 406), and the river formed the central backbone first of the kingdom of the Franks and then of the Car-

Boundary of Roman Gaul

olingian empire. When in 843 that empire was divided, stretches of the Rhine formed the eastern boundary of the central part, Lotharingia, until 870 when the Rhine again became the central axis of a political unit, the Holy Roman Empire. Subsequent events shifted the axis of this empire eastward and caused political disintegration along the Rhine. The Thirty Years' War (1618–48) ended with the final separation of the Rhine headwaters and delta area from Germany and a gradual advance of France toward the Rhine, which it reached under Louis XIV through his acquisition of Alsace.

The French Revolutionary Wars included further French advances, and the Treaty of Lunéville (1801) made the Rhine, along most of its course, France's eastern boundary. But France advanced beyond the Rhine and included northwestern Germany within its borders, and the Confederation of the Rhine, created by Napoleon, extended French control as far as the Elbe and Neisse rivers. The resultant upsurge of German nationalism was expressed by E.M. Arndt, who in 1813 wrote, "The Rhine is Germany's river, not its boundary." The Congress of Vienna, nevertheless, left France in possession of Alsace and thus with a Rhine frontier. Ambitions of Napoleon III to acquire further Rhenish territory strongly aroused German feelings. In 1840 Max Schneckenburger wrote his patriotic poem "Die Wacht am Rhein" ("The Watch on the Rhine"), which was set to music by Karl Wilhelm in 1854 and became the rousing tune of the Prussian armies in the Franco-German War of 1870–71. One result of this war was that France lost Alsace and thus its Rhine frontier, which it regained after World War I.

The fortified defensive system of the Maginot Line (built in 1927–36) adjoined the French bank of the upper Rhine from the Swiss frontier to near Lauterbourg. The opposing *Westwall,* or Siegfried Line (1936–39), adjoined the German bank from the Swiss frontier to near Karlsruhe.

Events after World War II suggested that the struggle for possession of the Rhine had been superseded by a trend toward economic and even political union of the riparian states. In addition, the increased pollution of the Rhine has resulted in growing international cooperation to combat the threat. (K.A.Si.)

RHÔNE RIVER

The Rhône, a historic river of Switzerland and France and one of the most significant waterways of Europe—it is the only major river flowing directly to the Mediterranean Sea—is thoroughly Alpine in character. In this respect it differs markedly from its northern neighbour, the Rhine, which leaves all of its Alpine characteristics behind when it leaves Switzerland. The scenic and often wild course of the Rhône, the characteristics of the water flowing in it, and the way it has been used by humans have all been shaped by the influences of the mountains, right down to the river mouth, where sediments marking the Rhône's birth in an Alpine glacier are carried into the warmer waters of the Mediterranean.

The Rhône is 505 miles (813 kilometres) long and has a drainage basin of some 37,750 square miles (97,775 square kilometres). The course of the river can be divided into three sectors lying, respectively, in the Alps, between the Alps and the Jura Mountains and through the latter, and finally in the topographical furrow of Alpine origin running from the city of Lyon to the sea.

Physical features. *Physiography.* The Rhône originates in the Swiss Alps, upstream from Lake Geneva. It comes into being at an altitude of about 6,000 feet (1,830 metres), emerging from the Rhône Glacier, which descends the south flank of the Dammastock, a nearly 12,000-foot peak. The river then traverses the Gletsch Basin, from which it escapes through a gorge, and flows along the floor of the Goms Valley at an altitude between 4,000 and 4,600 feet. It next enters another gorge before reaching the plain of the Valais, which extends between the towns of Brig and Martigny, and descends in altitude 2,300 to 1,600 feet. In crossing this high and rugged mountain area, the river makes successive use of two structural troughs. The first runs between the ancient crystalline rock massifs of the Aare and of the Gotthard; farther downstream the second runs between the arched rock mass of the Bernese Oberland and, on the south, the massive rock face of the Pennine Alps. From Brig onward, the landscape changes. During the last Ice Age a large glacier, fed by several small ones, plowed down the valley floor of the Valais, and, except for some harder rock obstacles found near the town of Sion, succeeded in widening and deepening the narrow valley floor. As it did so, it held back both the upper Rhône and those of its tributaries that come down from the Pennine Alps. When the ice sheets retreated, both the tributaries—the Vispa, Navigenze, Borgne, and Drance—and the Rhône cut new, deep gorges to connect their lower courses to the new valley floor. These gorges have created considerable difficulty for modern transportation, necessitating a series of hairpin-bend road links.

The Alpine sector

After Martigny, where the valley floor is wider, the youthful Rhône thrusts northward at a right angle, cutting across the Alps through a transverse valley. At first, near the town of Saint-Maurice, this is no more than a small gorge, but it soon becomes wider and flatter. There, too, the river route has been assisted by structural factors, specifically by a dip in the crystalline rock massifs running from Mont Blanc to the Aare and by the discontinuity between the limestone masses of the Dents du Midi and of the Dent de Morcles. Across the mountain barrier the muddy waters of the Rhône enter another wide plain surrounded by high mountains and then plunge into the clearer, stiller waters of Lake Geneva, forming an enlarging delta.

The second sector of the Rhône's course commences with Lake Geneva, large (224 square miles) and deep (1,000 feet) and lying between Switzerland and France in a basin hollowed out of the less resistant terrain by the former Rhône Glacier. Upon leaving Lake Geneva, which has turned the course of the river to the southwest and decanted the sediment from its waters, the Rhône very quickly regains in full the milky colour so characteristic of Alpine rivers. Just below the city of Geneva, it receives its powerful tributary the Arve, which rushes down from the glaciers of Mont Blanc.

The middle course

From its juncture with the Arve to the French city of Lyon, the Rhône has to cross a difficult obstacle, the undulating series of ridges forming the Jura Mountains. It does this by cutting through deep longitudinal valleys called *vaux* and transverse valleys called *cluses,* which were formed when the Jura Mountains were uplifted during the Alpine orogeny. As a result, the river follows a complicated zigzag course. At the town of Bellegarde the river is joined from the north by the Valserine and, swinging south, plunges into a deep gorge now submerged in the 14-mile-long Génissiat Reservoir. In the wider sections of its course in this region, the Rhône runs through glacier-excavated basins that its own deposits have barely filled, causing intermittent marshy areas. It is also joined by the Ain, from the north, and, on the left bank, by the Fier and Guiers. The river next widens, and the terrain becomes less hilly and, at Le Parc (some 95 miles above Lyon), becomes officially navigable, although the average depth is no more than three feet.

The lower sector

At Lyon the Rhône enters its third sector as it heads south toward the Mediterranean, which is characterized by the great north–south Alpine furrow that is also drained by its principal tributary, the Saône. The latter lies in the basins that the Ice Age glaciers hollowed out between the Jura Mountains to the east and, farther west, the eastern edge of the Paris Basin and the uplands of the Massif Central. It forms an important commercial link to the industrialized regions of northern France. From the city of Lyon onward, the river occupies the trough lying between the Massif Central and the Alps, a channel up which the sea of the Tertiary period (66.4 to 1.6 million years ago) ascended covering the present Rhône valley. A body of water, Lake Bresse, spread over the Saône basin. Into this lake drained a river—the present Rhine—which then flowed south through the valley and into the Saône basin; later tectonic movements caused the Rhine to reverse its flow, and the Doubs, a tributary of the Saône, now partly follows the former Rhine drainage pattern. In the late Tertiary the gulf of the sea was uplifted to expose the lower Rhône valley, and Lake Bresse drained out to the

south through the Saône River. Though the Rhône–Saône corridor is underlain by sediments laid down during the Tertiary period, much of its present surface is formed by debris deposited by valley glaciers that extended from the Alps during the Pleistocene epoch (1.6 million to 10,000 years ago). These sediments were instrumental in cutting deep channels through the edge of the crystalline Massif Central, as evidenced at Vienne and Tain. The valley consequently takes the form of a series of gorges and basins, the latter often having a series of terraces corresponding to variations in the levels of ice and of river. Although the tributaries—notably the Ardèche—rushing down into the Rhône from the Massif Central are formidable when in flood, the great Alpine rivers, the Isère, and the Durance, joining the left bank, are most important in their effect on riverbed deposits and on the volume of water. Below Mondragon the Rhône valley becomes wider, and what was once a marshy landscape open to flooding has been regulated by a series of dams and canals.

The river's delta begins near Arles and extends about 25 miles to the sea. Twin channels of the river, the Grand and Petit Rhône, enclose the Camargue region. This region, formed by alluvium, is continuously extending into the Mediterranean. The finer materials are carried by onshore currents to form the barrier beaches of the coast and the sandbars closing off the Étang de Berre. One part of the delta has been set aside for a nature reserve, thereby protecting the feeding and nesting grounds of flamingos, egrets, ibis, and other rare species. Since 1962 the left bank of Fos has been transformed into a vast industrial complex consisting of port facilities, refineries, oil-storage tanks, and steel mills.

Hydrology. The flow regime of the Rhône owes its remarkable mean volume to the influence of the Alps. At Lyon the flow amounts to 22,600 cubic feet (640 cubic metres) per second; there, the Saône alone contributes 14,100 cubic feet per second. The Isère adds another 12,400 cubic feet per second. The melting of the Alpine snows gives the highest mean flows in May, while the Saône attains its maximum in January. The flood volumes of spring and autumn are formidable, reaching 460,000 cubic feet per second for the Rhône at Beaucaire, just above the delta. Thus, the Rhône has an abundant flow but maintains a strong gradient almost to its mouth. At Lyon, for example, its altitude is 560 feet at 205 miles from the sea. As the size of the delta region testifies, the river transports enormous amounts of alluvial deposits and is also powerful enough to cut through a variety of rock masses. As a result, the Rhône of today is well adapted to the production of electricity and, though difficult to navigate in the past, is now an important waterway from the Mediterranean to Lyon.

Flood volumes

The economy. The Rhône basin constitutes one of the great economic regions of Switzerland and of France, draining rich plains as well as an important part of the Alps. The utilization of the region by humans, however, has required a long historical struggle, which entered a decisive phase only in the second half of the 20th century. The economy of the Rhône region consists of five major elements: agriculture, industry, energy, tourism, and transportation.

Agriculture in the Rhône valley largely covers the low areas, plains, and islands. In the canton of Valais, the Rhône has been diked and narrowed, and the surrounding plain has been drained. Comparable works have been carried out in France, notably on the Isère, at Combe de Savoie and Grésivaudan, and on the upper Durance. River waters are used extensively for irrigation. Forage crops and livestock raising coexist with vineyards, fruit orchards, and vegetable farming; the Camargue region is noted for its rice fields.

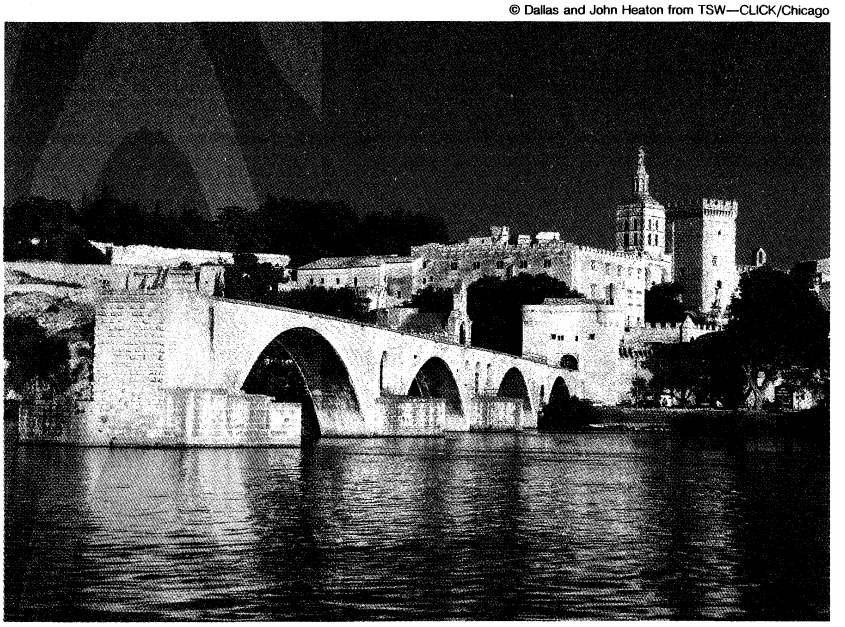
Industries, both large and small, have been established throughout the region. Notable concentrations include the aluminum and chemical plants in Valais, the oil refineries at Lyon, and the refineries and steel mills at Fos. The production of hydroelectricity is evident throughout the length of the Rhône and particularly so in its lower reaches, where a series of dam projects have harnessed more than half of the entire potential hydroelectric power of the river. In France several nuclear generating stations utilize river waters for cooling purposes.

The course of the Rhône has long attracted tourists, and tourism has played an increasingly key role in the regional economy since the mid-20th century. The great variety of recreational activities offered—from skiing and climbing in the Alps, to visiting the historic cities of Provence, to horseback riding in the Camargue—have been key to the river's popularity.

Navigation has always been carried on, particularly between Lyon and the sea, and the Rhône traditionally has been the transportation funnel between northern and southern France; the Rhône valley in Valais served a similar function in Switzerland, particularly with the construction of a number of rail and road tunnels under some of the mountain barriers. The most extensive improvements to the river itself have taken place between Lyon and the Mediterranean: shoals have been submerged under the reservoirs created by dams or bypassed by canals, and the original gradient has been replaced by a succession of level reaches and locks.

Transportation

© Dallas and John Heaton from TSW—CLICK/Chicago



The Rhône River at Avignon, Fr., with the Saint-Bénézet Bridge in the foreground and the Papal Palace in the background.

History. Great cities attest the antiquity and the strength of people's interest in the region, which long ago was influenced by Celtic settlement and then by Roman domination. Brig, Sion, and Martigny in the Alpine section; Lausanne and Nyon on Lake Geneva; Lyon, at one of the major European crossways; and the Provençal cities of Nîmes, Arles, and Orange all contain evidence of their Roman past. From AD 1033 much of the region was controlled by the house of Savoy; ultimately, Valais, Geneva, and Vaud joined the Swiss Confederation, while Savoy itself became part of France. During the 14th and early 15th centuries, Avignon (located just north of the Rhône delta) was the residence of the popes of the Avignon papacy and antipopes of the Western Schism. The river, once a spearhead for the penetration of Mediterranean cultures and peoples into northern Europe, again became a routeway for invasion, when Allied armies followed it north after landing in southern France during World War II. (P.V./A.Di.)

SEINE RIVER

The Seine River, 485 miles (780 kilometres) long, with its tributaries drains an area of about 30,400 square miles (78,700 square kilometres) in northern France; it is one of Europe's great historic rivers, and its drainage network carries most of the French inland waterway traffic. Since the early Middle Ages it has been above all the river of Paris, and the mutual interdependence of the river and the city that was established at its major crossing points has been indissolubly forged. The fertile centre of its basin in the Île-de-France was the cradle of the French monarchy and the nucleus of the expanding nation-state and is still its heartland and metropolitan region.

Physical features. *Physiography.* The Seine rises at 1,545 feet (471 metres) above sea level on the Mont Tasselot in the Côte d'Or region of Burgundy but is still only a small stream when it traverses porous limestone country beyond Châtillon. Flowing northwest from Burgundy, it enters Champagne above Troyes and traverses the dry chalk plateau of Champagne in a well-defined trench. Joined by the Aube near Romilly, the river bears west to skirt the Île-de-France in a wide valley to Montereau, where it receives the Yonne on its left bank. This tributary is exceptional in rising beyond the sedimentary rocks of the Paris Basin on the impermeable crystalline highland of the Morvan, a northward extension of the Massif Central. Turning northwest again, the Seine passes Melun and Corbeil as its trenched valley crosses the Île-de-France toward Paris. As it enters Paris, it is joined by its great tributary the Marne on the right, and, after traversing the metropolis, it receives the Oise, also on the right. In its passage through Paris, the river has been trained and narrowed between riverside quays. Flowing sluggishly in sweeping loops, the Seine passes below Mantes-la-Jolie across Normandy toward its estuary in the English Channel. The broad estuary opens rapidly and extends for 16 miles below Tancarville to Le Havre; it experiences the phenomenon of the tidal bore, which is known as the *mascaret,* although continued dredging since 1867 has deepened the river so that the *mascaret* has gradually diminished.

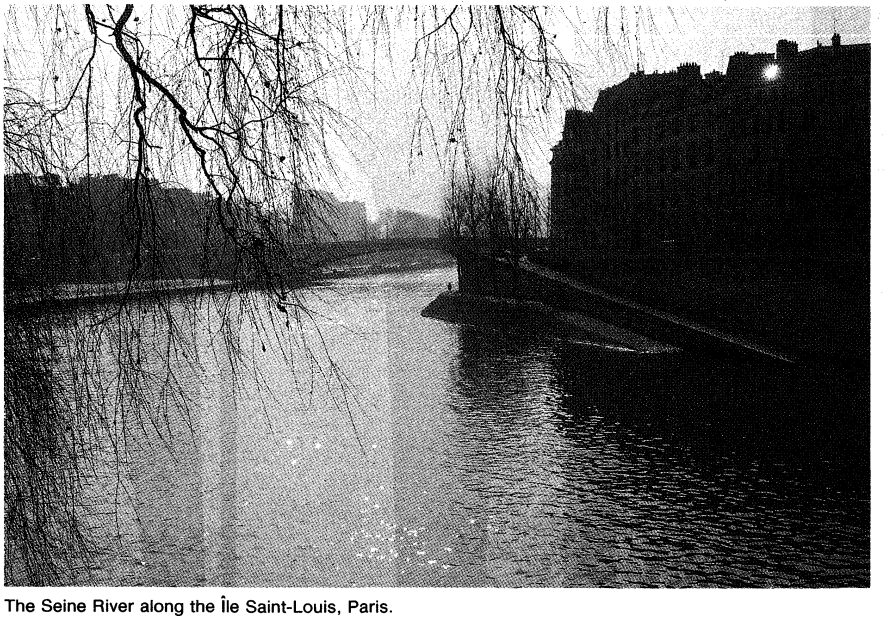
From its source to Paris, the Seine traverses concentric belts of successively younger sedimentary rocks, infilling a structural basin, the centre of which is occupied by the limestone platforms of the Île-de-France immediately surrounding Paris. The rocks of this basin are inclined gently toward Paris at the centre and present a series of outward-facing limestone (including chalk) escarpments (*côtes*) alternating with narrower clay vales. The *côtes* are breached by the Seine and its tributaries, which have made prominent gaps. As they converge upon Paris, the trenchlike river valleys separate a number of islandlike limestone platforms covered with fertile, easily worked windblown soil (*limon*). These platforms have provided rich cereal-growing land from time immemorial and constitute the Île-de-France. The lower course of the Seine, below Paris, is directed in a general northwesterly direction toward the sea, in conformity with the trend of the lines of structural weakness affecting the northern part of the basin. The English Channel breaches the symmetry of the basin on its northern side, interrupting the completeness of the concentric zones. Still in the chalk belt, the river enters the sea.

The geologic background

The basin of the Seine presents no striking relief contrasts. Within 30 miles of its source the river is already below 800 feet, and at Paris, 227 miles from its mouth, it is only 80 feet above sea level. It is thus slow flowing and eminently navigable, the more so because its regime is generally so regular.

Hydrology. Most of the river basin is formed of permeable rocks, the absorptive capacity of which mitigates the risk of river floods. Precipitation throughout the basin is modest, generally 25 to 30 inches (650 to 750 millimetres), and is evenly distributed over the year as rain, with snow infrequent except on the higher southern and eastern margins. The Yonne—unique among the tributaries in being derived from impermeable, crystalline highlands, where there is also considerable winter snow—also has the greatest influence on the Seine's regime (flow) because of the great variability of its flow; but the Seine is the most regular of the major rivers of France and the most naturally navigable. Occasionally the summer level is considerably reduced (such as in the summers of 1947 and 1949), but the sandbanks that are so typical of the

© Dana Hyde—Photo Researchers



The Seine River along the Île Saint-Louis, Paris.

Loire do not appear. Low water is further masked by the regularization of the river that has been carried out to improve its navigability. Winter floods are rarely dangerous, but in January 1910 exceptionally heavy rainfall caused the river to rise above 28 feet at Paris, flooding the extensive low-lying quarters along its ancient meander loop (the Marais). To match this high level it is necessary to go back to February 1658; but in January 1924 and also in January 1955 the river again rose to more than 23 feet in Paris. The average flow at Paris is about 10,000 cubic feet (280 cubic metres) per second, as compared with the 1910 flood rate of about 83,000 and the 1947 and 1949 minimums of about 700.

Economic aspects of the river

The economy. The Seine, especially below Paris, is a great traffic highway. It links Paris with the sea and the huge maritime port of Le Havre. Rouen, although some 75 miles from the sea, was France's main seaport in the 16th century, but it was surpassed by Le Havre in the 19th century. Vessels drawing up to 10 feet (3.2 metres) can reach the quays of Paris. Most of the traffic, which chiefly consists of heavy petroleum products and building materials, passes upstream to the main facilities of the port of Paris at Gennevilliers. The lower Seine system is connected with that of the Rhine by way of the Marne, and the Oise links it with the waterways of Belgium. The links with the Loire waterway and with the Saône-Rhône, dating from the 17th and 18th centuries when connecting canals were built, are now of minor importance. The water of the Seine is an important resource for the riverine population. Large electric power stations, both thermal and nuclear, draw their cooling water from the river. In addition, half of the water used in the region around Paris, both for industry and for human consumption, and three-fourths of the water used in the region between Rouen and Le Havre, is taken from the river.

Development of the river. Although the regime of the Seine is relatively moderate, improvements have been considered necessary since the beginning of the 19th century. To improve navigation, the water level was raised by means of dams and by storage reservoirs in the basin of the Yonne River. Lake Settons (1858), originally designed for the flotation of wood, and Crescent (1932) and Chaumeçon (1934) reservoirs have proved useful in reducing floods as well as in ensuring a constant water supply in summer. Upstream from the basin four large storage reservoirs have been built since 1950 on the Yonne, Marne, and Aube, as well as on the Seine itself. These relatively shallow impoundments (averaging about 25 feet in depth) cover large areas. The Seine Reservoir, for example, covers some 6,175 acres (2,500 hectares), while the Marne Reservoir, with an area of about 11,900 acres, is the largest artificial lake in western Europe. Surrounded by woodland and countryside, these reservoirs have become bird sanctuaries and tourist attractions in a new nature reserve.

(A.E.Sm./M.Da.)

Central European drainage systems

DANUBE RIVER

The Danube is the second longest river of Europe after the Volga. It rises in the Black Forest mountains of West Germany and flows for approximately 1,770 miles (2,850 kilometres) to its mouth on the Black Sea. Along its course, it passes through eight countries under six variations of its name. In West Germany and Austria it is known as the Donau, in Czechoslovakia as the Dunaj, in Hungary as the Duna, in Yugoslavia and Bulgaria as the Dunav, in Romania as the Dunărea, and in the Soviet Union as the Dunay.

The Danube played a vital role in the settlement and political evolution of central and southeastern Europe. Its banks, lined with castles and fortresses, formed the boundary between great empires, and its waters served as a vital commercial highway between nations. The river's majesty has long been celebrated in music. The famous waltz *An der schönen, blauen Donau* (1867; *The Blue Danube*), by Johann Strauss the Younger, became the symbol of imperial Vienna. In the 20th century the river has continued its role as an important trade artery. It has been harnessed for hydroelectric power, particularly along the upper courses, and the cities along its banks—including the national capitals of Vienna, Budapest, and Belgrade—have depended upon it for their economic growth.

Physical features. *Physiography.* The Danube's vast drainage of some 315,000 square miles (817,000 square kilometres) includes a variety of natural conditions that affect the origins and the regimes of its watercourses. They favour the formation of a branching, dense, deepwater river network that includes some 300 tributaries, more than 30 of which are navigable. The river basin expands unevenly along its length. It covers about 18,000 square miles at the Inn confluence, 81,000 square miles after joining with the Drava, and 228,000 square miles below the confluences of its most affluent tributaries, the Sava and the Tisza. In the lower course the basin's rate of growth decreases. More than half of the entire Danube basin is drained by its right-bank tributaries, which collect their waters from the Alps and other mountain areas and contribute up to two-thirds of the total river runoff or outfall.

Three sections of the Danube

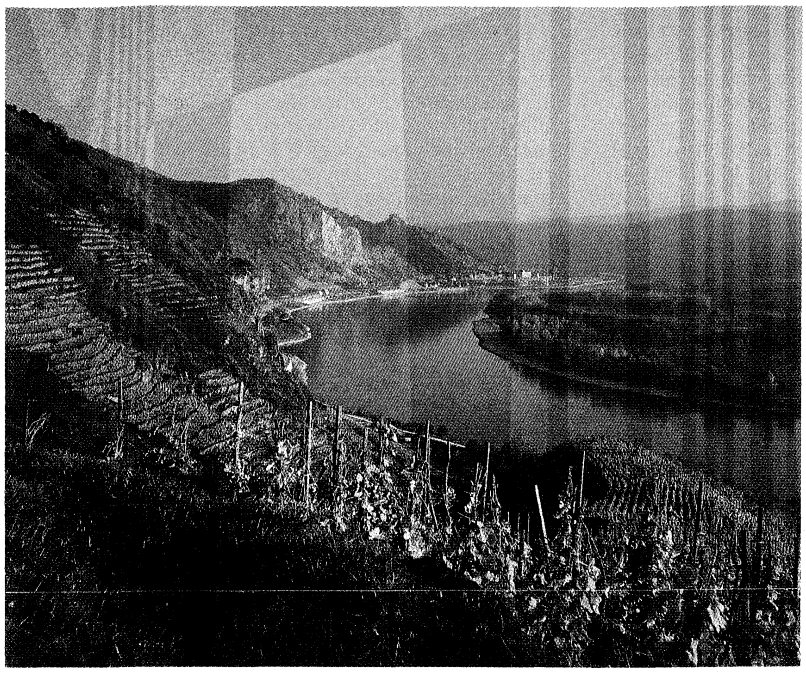
Three sections are discernible in the river's basin. The upper course stretches from its source to the gorge, called the Hungarian Gates, in the Austrian Alps and the Western Carpathian Mountains. The middle course runs from the Hungarian Gates to the Iron Gate Gorge in the Southern Romanian Carpathians. The lower course flows from the Iron Gate to the deltalike estuary at the Black Sea.

The upper Danube springs as two small streams—the Breg and Brigach—from the eastern slopes of the Black Forest mountains of West Germany, which partially consist of limestone. From Donaueschingen, where the headstreams unite, the Danube flows northeastward in a narrow, rocky bed. To the north rise the wooded slopes of the Swabian and the Franconian mountains; between Ingolstadt and Regensburg the river forms a scenic canyonlike valley. To the south of the river course stretches the large Bavarian Plateau, covered with thick layers of river deposits from the numerous Alpine tributaries. The bank is low and uniform, composed mainly of fields, peat, and marshland.

At Regensburg the Danube reaches its northernmost point, from which it veers south and crosses wide, fertile, and level country. Shortly before it reaches Passau on the Austrian border, the river narrows and its bottom abounds with reefs and shoals. The Danube then flows through Austrian territory, where it cuts into the slopes of the Bohemian Forest and forms a narrow valley. In order to improve navigation, dams and protecting dikes have been built near Passau, Linz, and Ardagger. The upper Danube, some 600 miles long, has a considerable average inclination of the riverbed (0.93 percent) and a rapid current of two to five miles per hour. Depths vary from three to 26 feet (one to eight metres). The Danube swells substantially at Passau where the Inn River, its largest upstream tributary, carries more water than the main river. Other major tributaries in the upper Danube course include the Iller, Lech, Isar, Traun, Enns, and Morava rivers.

In its middle course the Danube looks more like a flatland river, with low banks and a bed that reaches a width of more than one mile. Only in two sectors—at Visegrád (Hungary) and the Iron Gate—does the river flow through narrow, canyonlike gorges. The basin of the middle Danube exhibits two main features—the flatland of the Little Alfold and the Great Alfold plains, and the low peaks of the Western Carpathians and the Transdanubian Mountains.

The Danube enters the Little Alfold plain immediately after emerging from the Hungarian Gates Gorge near Bratislava, Czech. There the river stream slows down abruptly and loses its transporting capacity, so that enormous quantities of gravel and sand settle on the bottom. A principal result of this deposition has been the formation of two islands, one on the Czech side of the river and the other on the Hungarian side, which combined have an area of about 730 square miles that support some 190,000 inhabitants in more than 100 settlements. The silting hampers navigation and occasionally divides the river into two or more channels. East of Komárno the Danube enters the Visegrád Gorge, squeezed between the foothills of the Western Carpathian and the Hungarian Transdanu-



Vineyards along the Danube River in the Wachau region, Austria.
G. Hofmann—Superstock

bian Mountains. The steep right bank is crowned with fortresses, castles, and cathedrals of the Hungarian Árpád dynasty of the 10th to 15th century.

The Danube then flows past Budapest and across the vast Great Alfold plain until it reaches the Iron Gate gorge. The riverbed is shallow and marshy, and low terraces stretch along both banks. River accumulation has built a large number of islands, including Csepels Island near Budapest. In this long stretch the river takes on the waters of its major tributaries—the Drava, the Tisza, and the Sava—which create substantial changes in the river's regime. The average runoff increases from about 83,000 cubic feet (2,400 cubic metres) per second north of Budapest to 200,000 cubic feet at the Iron Gate. The river valley looks most imposing there, and the river's depth and current velocity fluctuate widely. The rapids and reefs of the Iron Gate once made the river unnavigable until a lateral navigation channel and a parallel railway allowed rivercraft to be towed upstream against the strong current.

The Iron Gate

Beyond the Iron Gate the lower Danube flows across a wide plain; the river becomes shallower and broader, and its current slows down. To the right, above steep banks, stretches the tableland of the Danubian Plain of Bulgaria. To the left lies the low Romanian Plain, which is separated from the main stream by a strip of lakes and swamps. The tributaries in this section are comparatively

The Danube River basin and its drainage network.

small and account for only a modest increase in the total runoff. They include the Olt, the Siret, and the Prut. The river is again obstructed by a number of islands. Just south of Cernavodă, the Danube heads northward until it reaches Galaţi, where it veers abruptly eastward. Near Tulcea, some 50 miles from the sea, the river begins to spread out into its delta.

The river splits into three channels—the Chilia, which carries 63 percent of the total runoff; the Sulina, which accounts for 16 percent; and the Sfîntu Gheorghe (St. George), which carries the remainder. Navigation is possible only by way of the Sulina Channel, which has been straightened and dredged along its 39-mile length. Between the channels, a maze of smaller creeks and lakes are separated by oblong strips of land called *grinduri*. Most *grinduri* are arable and cultivated, and some are overgrown with tall oak forests. A large quantity of reeds that grow in the shallow-water tracts are used in the manufacture of paper and textile fibres. The Danube delta covers an area of some 1,660 square miles and is a comparatively young formation. About 6,500 years ago the delta site was a shallow cove of the Black Sea coast, but it was gradually filled by river-borne silt; the delta continues to grow seaward at the rate of 80 to 100 feet annually.

Hydrology. The different physical features of the river basin affect the amount of water runoff in its three sections. In the upper Danube the runoff corresponds to that of the Alpine tributaries, where the maximum occurs in June when melting of snow and ice in the Alps is the most intensive. Runoff drops to its lowest point during the winter months.

In the middle basin the phases last up to four months, with two runoff peaks in June and April. The June peak stems from that of the upper course, reaching its maximum 10 to 15 days later. The April peak is local. It is caused by the addition of waters from the melting snow in the plains and from the early spring rains of the lowland and the low mountains of the area. Rainfall is important; the period of low water begins in October and reflects the dry spells of summer and autumn that are characteristic of the low plains. In the lower basin all Alpine traits disappear completely from the river regime. The runoff maximum occurs in April, and the low point extends to September and October.

The river carries considerable quantities of solid particles, nearly all of which consist of quartz grains. The constant shift of deposits in different parts of the riverbed forms shoals. In the stretches between Bratislava and Komárno and in the Sulina Channel, draglines are constantly at work to maintain the depth needed for navigation. The damming of the river has also changed the way in which sediments are transported and deposited. Water impounded by reservoirs generally loses its silt load, and the water flowing out of the dam—which is relatively silt-free—erodes banks farther downstream.

The temperature of the river waters depends on the climate of the various parts of the basin. In the upper course, where the summer waters derive from the Alpine snow and glaciers, the water temperature is low. In the middle and lower reaches summer temperatures vary between 71° and 75° F (22° and 24° C), while winter temperatures near the banks and on the surface drop below freezing. Upstream from Linz the Danube never freezes entirely because the current is turbulent. The middle and lower courses, however, become icebound during severe winters. Between December and March, periods of ice drift combine with the spring thaw, causing floating ice blocks to accumulate at the river islands, jamming the river's course, and often creating major floods.

The winter freeze

The natural regime of river runoff changes constantly as a result of the introduction of stream-regulating equipment, including dams and dikes. The mineral content of the river is greater during the winter than the summer. The content of organic matter is relatively low, but pollution increases as the waters flow past industrial areas. The river's chemistry also changes as city sewerage and agricultural runoff find their way into the river.

The economy. The Danube is of great economic importance to the eight countries that border it—the Soviet Union, Romania, Yugoslavia, Hungary, Bulgaria, Czechoslovakia, Austria, and West Germany—all of which variously use the river for freight transport, the generation of hydroelectricity, industrial and residential water supplies, irrigation, and fishing. The movement of freight is the most important economic use of the Danube, and such cities as Izmail, Ukranian S.S.R.; Galaţi and Brăila, Rom.; Ruse, Bulg.; Belgrade, Serbia; Budapest; Bratislava, Czech.; Vienna; and Regensburg, W. Ger., are among the major ports. Since World War II, navigation has been improved by dredging and by the construction of a series of canals, and river traffic has increased considerably. The most important canals—all elements in a continentwide scheme of connecting waterways—include the Danube–Black Sea Canal, which runs from Cernovadă, Rom., to the Black Sea and provides a more direct and easily navigable link, and the Main–Danube Canal, being built to link the Danube to the Rhine and thus to the North Sea.

The Danube has been tapped for power, mainly in its upper course. The process, however, has spread downstream. One of the largest hydroelectric projects—the Djerdap High Dam and the Iron Gate power station—was built jointly by Yugoslavia and Romania. Not only does the project produce hydroelectricity but it also makes navigable what was once one of the most difficult stretches on the river.

Dams and irrigation

Industrial use of Danube waters is made at Vienna, Budapest, Belgrade, and Ruse. The main irrigated areas are along the river in Czechoslovakia, Hungary, Yugoslavia, and Bulgaria. The river, however, has nearly become unfit for irrigation as well as for drinking water because of the tremendous increase in pollutants; pollution has also diminished the once-rich fishing grounds, although some of the fish have moved to side lakes and swamps.

History. During the 7th century BC, Greek sailors reached the lower Danube and sailed upstream, conducting a brisk trade. They were familiar with the whole of the river's lower course and named it the Ister. The Danube later served as the northern boundary of the vast Roman Empire and was called the Danuvius. A Roman fleet patrolled its waters, and the strongholds along its shores were the centres of settlements, among them Vindobona (later Vienna), Aquincum (later Budapest), Singidunum (later Belgrade), and Sexantaprista (later Ruse).

During the Middle Ages the old fortresses continued to play an important role, and new castles such as Werfenstein, built by Charlemagne in the 9th century, were erected. When the Ottoman Empire spread from southeastern to central Europe in the 15th century, the Turks relied upon the string of fortresses along the Danube for defense. The Habsburg dynasty recognized the navigational potential of the Danube. Maria Theresa, queen of Hungary and Bohemia from 1740 to 1780, founded a department to oversee river navigation, and in 1830 a riverboat made a first trip from Vienna to Budapest, possibly for trading purposes. This trip marked the end of the river's importance as a line of defense and the beginning of its use as a channel of trade.

Regulated navigation on the Danube has been the subject of a number of international agreements. In 1616 an Austro-Turkish treaty was signed in Belgrade under which the Austrians were granted the right to navigate the middle and lower Danube. In 1774, under the Treaty of Küçük Kaynarca, Russia was allowed to use the lower Danube. The Anglo-Austrian and the Russo-Austrian conventions of 1838 and 1840, respectively, promoted free navigation along the entire river, a principle that was more precisely formulated in the Treaty of Paris of 1856, which also set up the first Danubian Commission with the aim of supervising the river as an international waterway. In 1921 and 1923 final approval of the Danube River Statute was granted by Austria, Germany, Yugoslavia, Bulgaria, Romania, Great Britain, Italy, Belgium, Czechoslovakia, Hungary, and Greece. The international Danube Commission was thus established as an authoritative institution with wide powers, including its own flag, the right to levy taxes, and diplomatic immunity for its members. It controlled navigation from the town of Ulm to the Black Sea and kept navigational equipment in good repair.

International control

During World War II, free international navigation along the course of the river was interrupted by the hostilities, and a consensus concerning the resumption of navigation was not reached until the Danubian Convention of 1948. The new convention provided for the Danubian countries alone to participate in a reconstituted Danube Commission; of these countries, only West Germany did not join the convention. (P.G.P.)

ELBE RIVER

The Elbe (Czech: Labe), one of the major waterways of central Europe, runs from Czechoslovakia through East and West Germany to the North Sea, flowing generally to the northwest. It rises on the southern side of the Krkonoše (Giant) Mountains near the border of Czechoslovakia and Poland. The river makes a wide arc across Bohemia (northwestern Czechoslovakia) and enters East Germany near Dresden. Between Wittenberge, E.Ger., and Lauenburg, W.Ger., it forms the boundary between the two German states. Above Hamburg the Elbe (now wholly in West Germany) splits into two branches; these rejoin farther downstream and the river then broadens into its estuary, the mouth of which is at Cuxhaven, where it flows into the North Sea.

The total length of the Elbe is 724 miles (1,165 kilometres), of which 352 miles are in East Germany. Its total drainage area is 55,620 square miles (144,060 square kilometres). Major tributaries are the Vltava (Moldau), Ohře (Eger), Mulde, and Saale, all of which join it from the left; and the Iser, Schwarze ("Black") Elster, Havel, and Alster from the right.

Physical features. *Physiography.* The Elbe is formed by the confluence of numerous headwater streams in the Krkonoše Mountains a few miles from the Polish-Czechoslovak frontier. It flows south and west, forming a wide arc for about 225 miles in Czechoslovakia to its confluence with the Vltava at Mělník and is joined 18 miles downstream by the Ohře. It then cuts to the northwest through the picturesque Elbe Sandstone Mountains, and, in a gorge four miles long, it enters East Germany. Between Dresden and Magdeburg the Elbe receives many long tributaries, of which all except the Schwarze Elster are left-bank streams. These are the Mulde and the Saale and its tributaries—including the Weisse ("White") Elster, the Unstrut, and the Ilm. These left-bank tributaries rise in the Ore Mountains or the Thuringian Forest and form the drainage basin of the middle Elbe with its geographic foci in Halle and Leipzig. Halle is on the Saale, just below the confluence of the Weisse Elster; Leipzig lies at the confluence of the Pleisse and the Weisse Elster. Below Magdeburg the Elbe receives most of its water from its right bank. Most of these tributaries rise in the uplands of Mecklenburg.

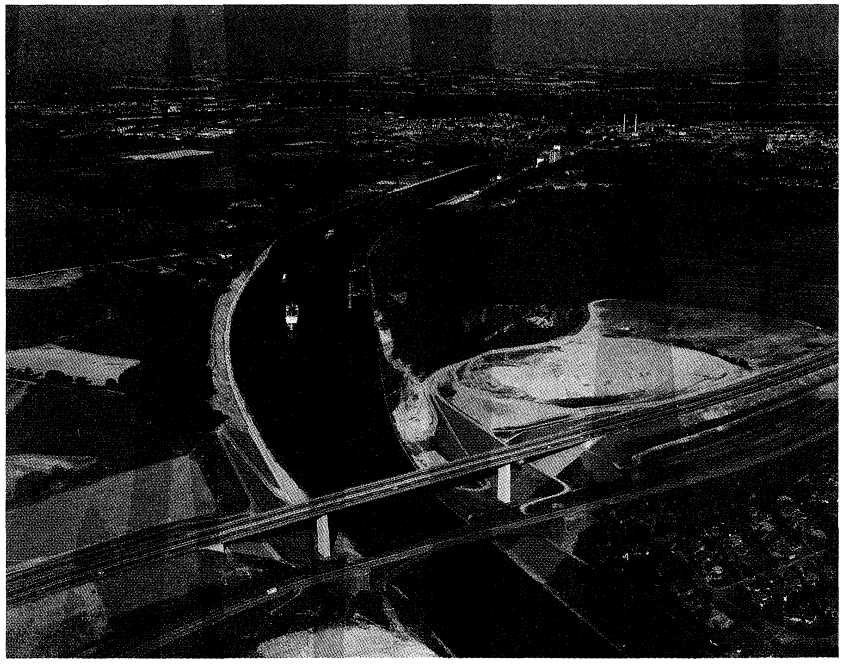
The middle Elbe

The river enters the North German Plain at Riesa, 25 miles below Dresden; below Riesa it meanders in a wide floodplain and has some abandoned loops. Dikes begin there and continue as far as the confluence of the Mulde. Between Wittenberg and Dessau the east–west valley floor narrows to five miles in width, and hilly land rises to the north (the Fläming Heath) and south. From Dessau to Magdeburg the floodplain widens, and dikes have been constructed continuously down to the sea. In its course below Magdeburg the floodplain is two miles wide down to the confluence with the Havel. The river keeps to the left of its floodplain and sometimes cuts into the low hills on its banks. Below the confluence with the Havel the river flows southeast–northwest; the floodplain widens and has distributaries and backwaters often flanked by low sandy hills (geest). Reclaimed salt marshes begin at Lauenburg. Above Hamburg—which the Elbe transverses in two arms, the Norder Elbe and the Süder Elbe—the floodplain is eight miles wide but narrows to four miles between the sandy geest of Schleswig-Holstein and the Lüneburg Heath.

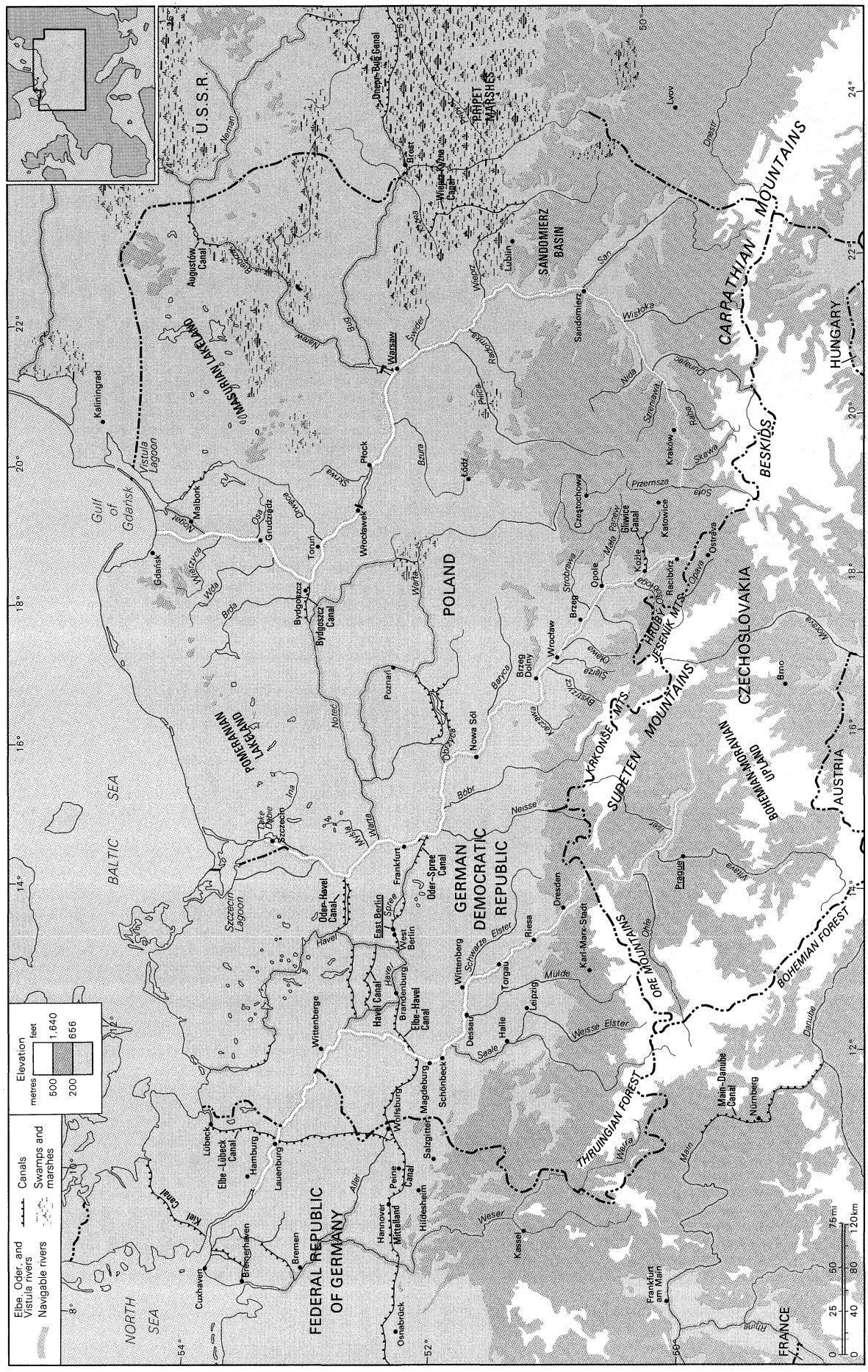
The estuary proper of the Elbe (Unterelbe) extends from Hamburg to Cuxhaven, a distance of about 55 miles. It varies in width from one to two miles, but much of it is occupied by mud flats and sandbanks. The main channel is buoyed and dredged. At high tide the channel has a depth of some 53 feet (16 metres). The south or left bank is low and marshy and the river has sandbanks; the right bank is steep below Hamburg, but farther downstream there are marshes, diked and drained, that are intensively cultivated. The great port city of Hamburg grew up on the Alster River on low sandy hills above the marshes. The modern port facilities have spread to the low-lying south bank of the Elbe.

Hydrology. The flow of water in the Elbe varies considerably with the amount of precipitation and thawing in its drainage basin. At Dresden the discharge rate averaged 11,200 cubic feet (317 cubic metres) per second in the period 1931–75, but the rate varied from a minimum of 800 cubic feet to a maximum of 118,700. At Neu-Darchau, about 140 miles above the mouth, the discharge rate was 24,700 cubic feet per second in the period 1926–65, with extremes of 5,100 and 127,700. These great variations sometimes hinder navigation. Although there are dams on the upper Elbe in Czechoslovakia and at Geesthacht,

A.G.E. FotoStock



The Kiel Canal, which runs from the mouth of the Elbe River to the Baltic Sea, at Kiel, W.Ger.



The Elbe, Oder, and Vistula river basins and their drainage network.

W.Ger., and large dams have been built on the Vltava and on the Saale in the Thuringian Forest, these are not sufficient to control the water level of the Elbe.

The lower course of the Elbe is tidal as far as the dam at Geesthacht, above Hamburg, where the river flow periodically reverses its direction. The average tide at Hamburg is about eight feet, and during storms the water may rise much higher, occasionally even flooding parts of the city.

Traffic on the Elbe

The economy. By means of the Elbe and its connecting waterways, vessels from Hamburg can navigate to Berlin, the central and southern sections of East Germany, and Czechoslovakia. The Mittelland Canal, a short distance below Magdeburg, runs westward about 200 miles to the Dortmund–Ems Canal, carrying barges of up to 1,000 tons to the West German industrial cities of Osnabrück, Hannover, Salzgitter, Hildesheim, Peine, and Wolfsburg and connecting with the Weser and Rhine rivers. The Elbe–Havel Canal carries traffic from Magdeburg eastward to the network of waterways around Berlin and farther on to Poland. The Kiel Canal runs from the mouth of the Elbe to the Baltic Sea, and the Elbe–Lübeck Canal, starting at Lauenburg, also runs to the Baltic, following an older (14th-century) canal. Another canal connects the lower Elbe with Bremerhaven on the Weser River. The Elbe itself is navigable for 1,000-ton barges as far as Prague through the Vltava. In East Germany it serves the river ports of Magdeburg, Schönebeck, Aken, Dessau, Torgau, Riesa, and Dresden, carrying bituminous coal, lignite, coke, metal, potash, grain, and piece goods. Although Hamburg lies far upstream from the mouth of the Elbe, it is one of the largest seaports in Europe; a six-line railway tunnel and a multilane road tunnel under the Elbe there are important links in trans-European traffic flows.

History. The basin of the Elbe has been settled since prehistoric times. Until the Middle Ages the river was the western boundary of the area inhabited by the northern Slavs. In the 12th century the Germans began to colonize the lands east of the Elbe and along the Baltic Sea. In World War II a point on the Elbe, near Torgau, was the meeting place of the U.S. and Soviet armies. Later the river formed part of the demarcation between East and West Germany.

Hamburg

The city of Hamburg dates from the early 9th century AD. Together with Lübeck, Hamburg established the Hanseatic League in 1241. Today it is West Germany's largest city, excluding West Berlin. Another ancient city on the Elbe is Magdeburg, which in the early 9th century was a trading post on the border between the Germans and the Slavs. In the 13th century it was a flourishing commercial city and an important member of the Hanseatic League. Today it is the largest inland harbour of East Germany. The other chief city of the Elbe is Dresden, founded about 1200. During the 18th century Dresden developed into a great centre of the fine arts, known as "Florence on the Elbe." Its beautiful architecture, almost completely destroyed during World War II, has been partially rebuilt. Other towns of historical interest along the Elbe include Wittenberg, the birthplace of the Protestant Reformation, and Meissen, which became famous for the manufacture of porcelain. (H.F./F.G.)

ODER RIVER

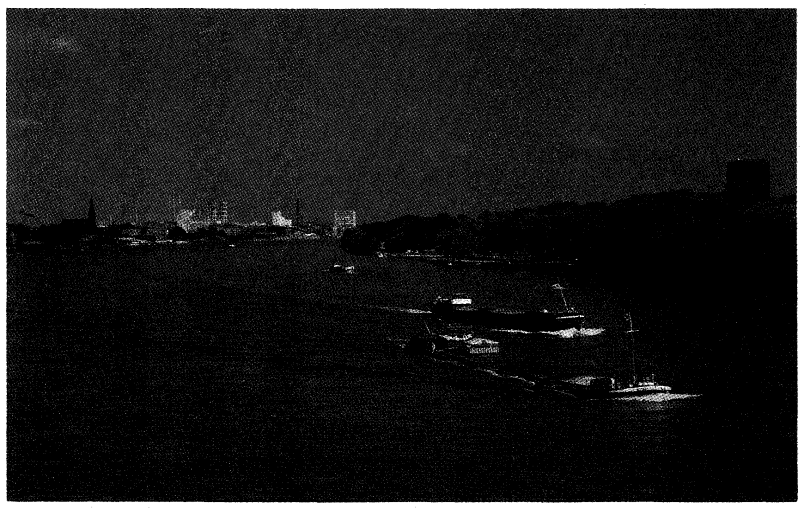
The Oder River, a vital economic artery in east central Europe, runs through the western portions of Poland and has considerable contemporary regional importance. It is one of the most significant rivers in the catchment basin of the Baltic Sea, second only to the Vistula in discharge and length. For the first 70 miles (112 kilometres) from its source, it passes through Czechoslovakia. For a distance of 116 miles in its middle reach, it constitutes the boundary between Poland and East Germany before reaching the Baltic Sea via a lagoon north of the Polish city of Szczecin. Called the Odra in Polish and Czech and the Oder in German, the river is an important waterway, navigable throughout most of its length. It forms a link, by way of the Gliwice Canal, between the great industrialized areas of Silesia (Śląsk), in southwestern Poland, and the trade routes of the Baltic Sea and beyond. The Oder is connected with the Vistula, Poland's largest river, by means of a water route utilizing the Warta and Noteć rivers, together with the Bydgoszcz Canal, and is tied in with the waterway system of western Europe by way of the Oder–Spree and Oder–Havel canals in East Germany.

The total length of the Oder River is 531 miles (854 kilometres), 461 miles of which lie in Poland. The total watershed area has been calculated at 46,000 square miles (119,000 square kilometres), of which about 90 percent is in Polish territory. The mean elevation of the Oder basin is 535 feet (163 metres) above sea level. From the river's source and over the greater part of its course, the Oder flows in a generally southeast–northwest direction; only from the junction with the Neisse (Polish: Nysa Łużycka) River does the northward trend toward the Baltic commence. The principal left-bank tributaries are the Opava of Czechoslovakia and the Osobłoga, Nysa Kłodzka, Oława, Ślęza, Bystrzyca, Kaczawa, Bóbr, and Neisse of Poland; from the east the main tributaries are the Olše of Czechoslovakia and the Kłodnica, Mała Panew, Strobrawa, Widawa, Barycz, Obrzyca, Warta, Myśla, and Ina of Poland. From the junction with the Opava the Oder is navigable for a distance of some 475 miles for 220 to 230 days of the year. Towns of particular importance along the Oder are Ostrava in Czechoslovakia, Frankfurt in East Germany, and Racibórz, Opole, Brzeg, Wrocław, Nowa Sól, and Szczecin in Poland.

Source

Physical features. *Physiography.* The Oder starts its course in Czechoslovakia, at an altitude of nearly 2,100 feet in the Hrubý Jeseník Mountains. Initially it runs as a mountain stream with a steep gradient that progressively lessens until the river reaches the floor of the structural depression called the Moravian Gate; from there the Oder continues its course in a wide valley. After receiving the Olše River, the Oder enters Poland and makes its way as a river that in a characteristic manner alternates between following ancient east–west stream valleys of glacial origin and crossing gaps cut in the intervening uplands. Where the Oder takes advantage of these preexisting valleys, it reaches widths as great as six miles or more, while in gaps it narrows to about a mile. Near Koźle the Gliwice Canal enters the Oder; and from there as far as Brzeg Dolny, a short way downstream from Wrocław, the river has a navigable channel controlled by locks. From Brzeg Dolny downstream until the final outflow into the Szczecin Lagoon, the river channel is fully improved. Beginning with the confluence with the Neisse River and continuing to just above Szczecin, the Oder becomes the borderline between Poland and East Germany. In this part of the valley the Oder–Spree and the Oder–Havel canals branch off to the west. Farther downstream the Oder valley contains numerous cross branches and parallel channels. About 50 miles from its outflow into the Baltic, the Oder splits into two main branches; the left canalized branch, called the Western Oder, passes through Szczecin and enters the Szczecin Lagoon directly, while the right branch, the Eastern Oder (in its final section called the Regalica), passes east of Szczecin via the large Lake Dąbie and then also enters the Szczecin Lagoon.

Hydrology. The Oder has a limited flow volume; its mean ratio of outflow to precipitation is the lowest among the rivers flowing into the Baltic. During low-water periods, in summer and autumn, the river is fed from storage reservoirs built in the upper tributaries. The mean water depth in the Oder Channel is three feet, and the mean velocity is three feet per second. In summer the upper reaches of the Oder system are flooded by heavy precipitation, while in spring the middle and lower reaches suffer from meltwater floods. Flow volume varies with the amount of precipitation. In the period 1951–80, for example, the discharge rate of the Oder's upper course averaged 1,560 cubic feet (145 cubic metres) per second, with extremes of 150 and 31,430; during that same period in the middle course the average was 18,820 cubic feet per second, with extremes of 5,510 and 76,630. The ice cover on the river lasts up to 40 days per year. As is the case with many of the world's great rivers flowing through heavily industrialized regions, the Oder's waters have become heavily polluted; of the fish that are still found in the river, the most common are bream and eel.



Barge traffic on the Oder River at Wrocław, Pol.
Ginette Laborde, Paris Charenton, Fr.

Improvements

The first hydraulic works—embankments and other structures for flood prevention—were started in the Oder valley as early as the 12th century; spillway dams built in the 13th century were in operation until the 18th century, when work was initiated on channel straightening by means of excavated cuts. Improvement of the straightened part of the Oder Channel was for the most part completed around 1900 (although final improvements were not made until after World War II), while control works in the middle and lower reaches were carried out in the interwar period.

The economy. The Oder River is an important element in the Polish economy, serving as a supplement to the heavily overburdened railway and highway systems linking the highly industrialized regions of the south with the largest Polish seaport, Szczecin, at the Oder's Baltic mouth. The river carries about 10 percent of the total tonnage handled by the port. The Oder is also used by the barges of East Germany, which travel over a system of navigable canals that connect the Oder with the central European waterway network. A system of navigable canals connects the Oder with the Vistula, Poland's largest river, and also with the rivers of the eastern portion of the country and the waterway system of the Soviet Union. This creates the possibility that the entire system may evolve into an all-water commercial route, transporting commodities from west to east and from east to west.

History. Because of its geographic situation, the Oder was, in ancient times, of major importance as the zone where people inhabiting southern and northern Europe came into contact with each other and exchanged cultural values. The first agricultural population arrived from the south after passing the Moravian Gate, which separates the Sudeten ranges from the Carpathian Mountains. Along the middle reach of the Oder there developed the pre-Lusatian and the Lusatian cultures (of the Bronze Age), which greatly affected the later evolution of the Slav population. In the area surrounding the Oder estuary, there was a mutual interpenetration by Scandinavian, Germanic, and Slav cultures. Finally, in the 9th and 10th centuries, the Polish state developed between the Oder and the Vistula. In the 13th century the German expansion dislodged Poland eastward, away from the Oder basin. But, on the basis of the 1945 Potsdam Conference between the Soviet Union, the United States, and Great Britain, the Polish nation returned to its former lands bordering the Oder River.

VISTULA RIVER

The Vistula (Polish: Wisła) is the largest river of Poland and of the drainage basin of the Baltic Sea. With a length of 651 miles (1,047 kilometres) and a drainage basin of some 75,100 square miles (194,500 square kilometres), it is a waterway of great importance to the nations of eastern Europe; more than 85 percent of the river's drainage basin, however, lies in Polish territory. The Vistula is connected with the Oder drainage area by the Bydgoszcz Canal. Eastward the Narew and Bug rivers and the Dnepr–Bug Canal link it with the vast inland waterway system of the Soviet Union.

Source

The source of the Vistula is found about 15 miles south of Bielsko-Biała on the northern slopes of the western Beskid range, in southern Poland, at an altitude of 3,629 feet (1,106 metres). It flows generally from south to north, through the mountains and foothills of southern Poland, across the lowland areas of the great North European Plain, ending in a delta estuary that enters the Baltic Sea near the port of Gdańsk. The average elevation of the Vistula basin is 590 feet above sea level; the mean river gradient is 0.10 percent, and the mean velocity in the river channel amounts to 2.6 feet per second. In addition to Poland's capital city, Warsaw, a number of large towns and industrial centres lie on the banks of the Vistula. These include Kraków, which was Poland's capital from the 11th century to the close of the 16th, Nowa Huta, Sandomierz, Płock, Toruń, Malbork, and Gdańsk. Numerous centres of tourism and recreation as well as many health resorts flank the Vistula valley. Here and there along the river rise the ruins of medieval strongholds, some of which have been restored.

Physical features. *Geology.* The present spatial pattern of the Vistula's tributary system and delta is the result of the changes in relief that occurred during the second half of the Tertiary and the Quaternary periods—*i.e.,* since about 30 million years ago. In the mountains the Vistula valley assumed its present shape much earlier and still reveals the way in which it has adapted itself to the geologic structure; in the lowland, on the other hand, the valley evolution was contingent on the history of the successive glaciations and, in particular, on changes during the interglacial period in which the Vistula abandoned its previous west–east valley and established its present northward course. The terminal part of the lower Vistula was finally stabilized in postglacial times, after the formation of the Baltic Sea.

A characteristic feature of the Vistula drainage basin is its asymmetry, with a predominance of right-bank over left-bank tributaries. This is the result of the general slant of the North European Plain in a northwesterly direction, which enabled the more powerful rivers of the Baltic drainage area to intercept the glacial streams flowing farther east.

Physiography. The course of the Vistula consists of three principal sections delineated by the San and Narew rivers, the two most prominent tributaries. The upper reach extends from the source to where the San joins its parent river near Sandomierz; its length is about 240 miles. The middle reach, from the mouth of the San to that of the Narew, below Warsaw, is about 170 miles long. Finally, the lower reach, extending to the Baltic, covers 240 miles from the mouth of the Narew to the mouth of the estuary into the Gulf of Gdańsk.

The upper course

In its upper course the Vistula is a mountain stream with a steep gradient of up to 5 percent. Its main sources are the Czarna Wisełka and the Biała Wisełka, two brooks that meet to form the Mała Wisła ("Small Vistula"), which then flows northward. Some 25 miles farther on, the river gradient decreases suddenly to some 0.04 percent; from there, after turning eastward, the Vistula enters Lake Goczałkowice, an artificial storage basin built in 1955. Upon exiting the lake, the Vistula assumes the character of a lowland stream, with its gradient decreasing to 0.03–0.02 percent in the middle reaches and to 0.02–0.002 percent in its final stages. At a distance of 65 miles from the source, the Vistula is joined by the Przemsza River, a left-bank tributary, after which—for 585 miles—it is navigable. After the Soła and Skawa—two right-bank tributaries—join the river, the Volga forces its way through a gap curved through a range of hills just before the city of Kraków. Channel improvements to this section have deprived the Vistula of much of its original character: several spillway steps have been constructed, creating a channel navigable by 300-ton barges. After passing through Kraków the Vistula turns to the east and, later, northeastward, crossing the wide Sandomierz Basin, where the valley is entered successively by the left-bank tributaries Szreniawa, Nida, Czarna, and Koprzywianka and from the right by the rivers Raba, Dunajec, Wisłoka, and San.

The inflow of the San River marks the beginning of the middle reaches of the Vistula, which then turns northward, breaching another gap through an upland area. In the course of its middle reaches the Vistula absorbs, from the left, the Radomka and Pilica and, from the right, the rivers Wieprz, Wilga, Świder, and Narew. Below the confluence with the Narew, where the lower reach of the river starts, the Vistula turns first to the west, and then after receiving the Bzura, a left-bank tributary, in a northwesterly direction; meanwhile from the right, the Skrwa and Drwęca join the river. In part of the valley, from the mouth of the Wieprz River to Toruń, the natural, untamed character of the Vistula predominates.

There the river runs in a channel 2,000 to 4,000 feet wide, practically devoid of controlling structures; in parts, the valley reaches widths up to six to nine miles, with the banks often 200 to 330 feet high. The low gradient of the river channel and abundant sandbanks render navigation difficult; in spring, when the ice cover breaks up and floats downstream, dangerous ice dams may form, causing the flooding of surrounding areas and often destroying embankments and bridges. A spillway step constructed at Włoclawek in 1968 initiated a series of improvements that continued through the 1980s.

From Toruń to its entry into the Baltic, the Vistula has been turned into a fully improved waterway. The 19th-century Bydgoszcz Canal, following an ancient glacial valley, links the Vistula with the Oder, the second largest of Polish rivers. Also near Bydgoszcz, the Vistula, having received a left-bank tributary in the Brda, turns northeastward in its third gap section cut through the Pomeranian highlands. Above Grudziądz the river finally turns northward to approach the Baltic. After receiving three further tributaries—the Osa from the right and the Wda and the Wierzyca from the left—the Vistula enters Żuławy Wiślane, its delta area, renowned for its splendidly fertile soils. Żuławy is a forestless plain, partly below sea level, threaded by the Vistula and its branches, together with a great number of canals and drainage ditches. Some of the local embankments and dikes date to the 13th century. During World War II a great part of Żuławy was flooded, but improvements were made in the postwar years.

The delta region

In the past the Vistula crossed its delta and entered the sea by two or more branch channels, notably the Nogat, which issued into the Vistula Lagoon, and the Leniwka (now called the Martwa Vistula), which followed the true Vistula Channel to the Gulf of Gdańsk. Improvements, the ultimate aim of which was to control the Vistula's outlet to the sea and make the entire delta region economically productive, were initiated at the end of the 19th century: first, a cut toward the open sea was excavated near Świbno to facilitate floodwater runoff and the removal of debris and ice carried by the river; later, all lateral watercourses were separated by locks, rendering them navigable, with controlled flows; the Świbno cut was extended into the open sea by lengthening the controlling embankments. This last change was intended to prevent the accumulation, at the river mouth, of the more than two million tons of sediment carried down annually by the Vistula.

The water level

Hydrology. Climatic variations in the Vistula basin cause a diversity in runoff and hence marked oscillations in the water level of the river, which averages 12 feet in the upper, 25 feet in the middle, and up to 33 feet in the lower reaches. Protracted low-water periods, lasting from late summer well into spring, are frequent. These hamper or entirely interrupt navigation. Spring floods caused by melting snow and ice in the whole drainage basin and summer floods resulting from heavy rains in the foothill and mountain regions are a common feature. During the period 1951–80 the mean flow of the upper course of the Vistula averaged about 2,200 cubic feet (62 cubic metres) per second, with extremes of 410 and 52,620 feet per second; the average for the middle course was about 20,900 feet per second, with extremes of 3,810 and 199,530; and the average for the lower course was 38,500 feet per second, with extremes of 8,940 and 276,870. Exceptionally heavy floods occurred in 1924, 1934, 1947, 1960, 1962, and 1970. There are a number of storage reservoirs in

D.C. Williamson, London

The Vistula River at Warsaw. In the background is the Old Town.

the valleys of the mountain tributaries that are intended to counteract excessive floods. Some newer, larger storage basins have been built.

Usually ice forms on the surface of the Vistula in the first half of January, breaking up toward the end of February. In the upper and lower reaches the duration of the ice sheet is from 20 to 40 days, in the middle reach it is 40 to 60 days, and in the estuary section, up to 20 days.

The quality of the Vistula's waters is affected by water-management structures such as dams and hydroelectric plants, by the discharge of municipal and industrial wastewater, and by agricultural and storm runoff. Although the upper reaches of the river remain relatively pure, the lower portions of the Vistula, in common with similar stretches of many of the great rivers of the world, exhibit a high degree of pollution.

Temperature regime

The mean annual temperature of the Vistula water is 46° F (8° C) in the upper reaches and 49° F (9° C) in the middle and lower reaches; in the middle and lower parts of the river the water is some 4° F (2° C) warmer than the mean annual air temperature of Poland. In winter the water temperature is 36° to 37° F (2° to 3° C); in summer it varies from 54° to 59° F (12° to 15° C). In river sections that are thermally affected by nearby industries, however, as in the regions of Kraków, Warsaw, and Włocławek, the water temperature is apt to be as much as 11° to 18° F (6° to 10° C) or even higher.

Plant and animal life. Higher-growth aquatic plant species most often encountered in the Vistula valley are, among plants submerged in the water, arrowhead (*Sagittaria sagittifolia,* variety *vallisinfolia*); among plants with floating leaves, the water lily (*Nuphar luteum*); and, among air-growing plants, sweet flag (*Acorus calamus*).

More than 40 kinds of fish exist in the Vistula. In the upper reach, turbot is the most common, with bream in the middle and lower reaches, and, in the waters of the estuary, salmon trout and vimba vimba. Species penetrating the river from the Baltic are found only sporadically.

The economy. The Vistula is connected with the Oder River by the Brda River, the Bydgoszcz Canal, and the Notec and Warta rivers; and in 1960 the Soviet Union, East Germany, and Poland agreed to establish permanent shipping lines along this route. In 1963 a canal was opened to avoid the natural hazards at the confluence of the Vistula and the Narew, improving the links between the Vistula and the Soviet waterways system.

Despite the Vistula's potential role as a transport link between the heavy industrial centres of southern Poland and the Baltic ports, navigational hazards have restricted its traffic. Nevertheless, attracted by water supply and by the possibilities of cheap transport rates for bulk materials, a number of large industrial projects have sprung up along the Vistula.

History. The Vistula played a prominent part in the ancient history of Poland. Since early Stone Age times the river served both as a trade route and as a means of expansion, from both north and south, for various peoples and cultures. Initially, raw materials and flint tools journeyed northward, while amber was sent to the south. By the time of the Roman Empire, the Vistula was one of the principal trade routes leading into central Europe; and from this period date the first historical references—by the classical geographers Pliny, Tacitus, and Ptolemy—to the Vistula and the Slav tribes living along its banks.

Medieval trade

Much later, in the early period of the Polish state (10th–13th century), the most important goods shipped over the Vistula route were salt, timber, grain, and building stone. The most intensive development of the Vistula as a trade route came from the 15th to 18th century, during which period a variety of hydraulic structures were put up, as well as embankments to provide flood protection. Many granaries and storehouses, built in the 14th century, line the banks of the Vistula. At the end of the 18th century, the partition of Poland between Prussia, Austria, and Russia put an end to the economic importance of the Vistula. Minor navigation improvements were undertaken only locally, in Prussia and in Austria. The major 19th-century improvements in the region of the delta and the construction of the Bydgoszcz Canal have been mentioned above. From 1920 to 1939 very little was done to improve the river channel. It was only after World War II that concerted efforts were undertaken to restore the Vistula to its historic function as a navigable waterway. This was done by the construction of a number of storage reservoirs and spillway dams in the river and its tributaries: the purpose was to take advantage of the river's hydroelectric potential and, at the same time, to adapt the channel to the travel of freight barges of 600- to 1,000-ton capacity.

A number of institutions are concerned with research on the Vistula and with keeping the waterway in operation. The highest authority coordinating activities in the field of research and deciding on technical expenditures and on navigational improvements is the Ministry of Environment Protection and Natural Resources. In addition, hydrologic measurements and investigations as well as engineering studies are carried out by the Institute for Meteorology and Water Management. (W.Pa./Je.P.)

Eastern European drainage systems

DNEPR RIVER

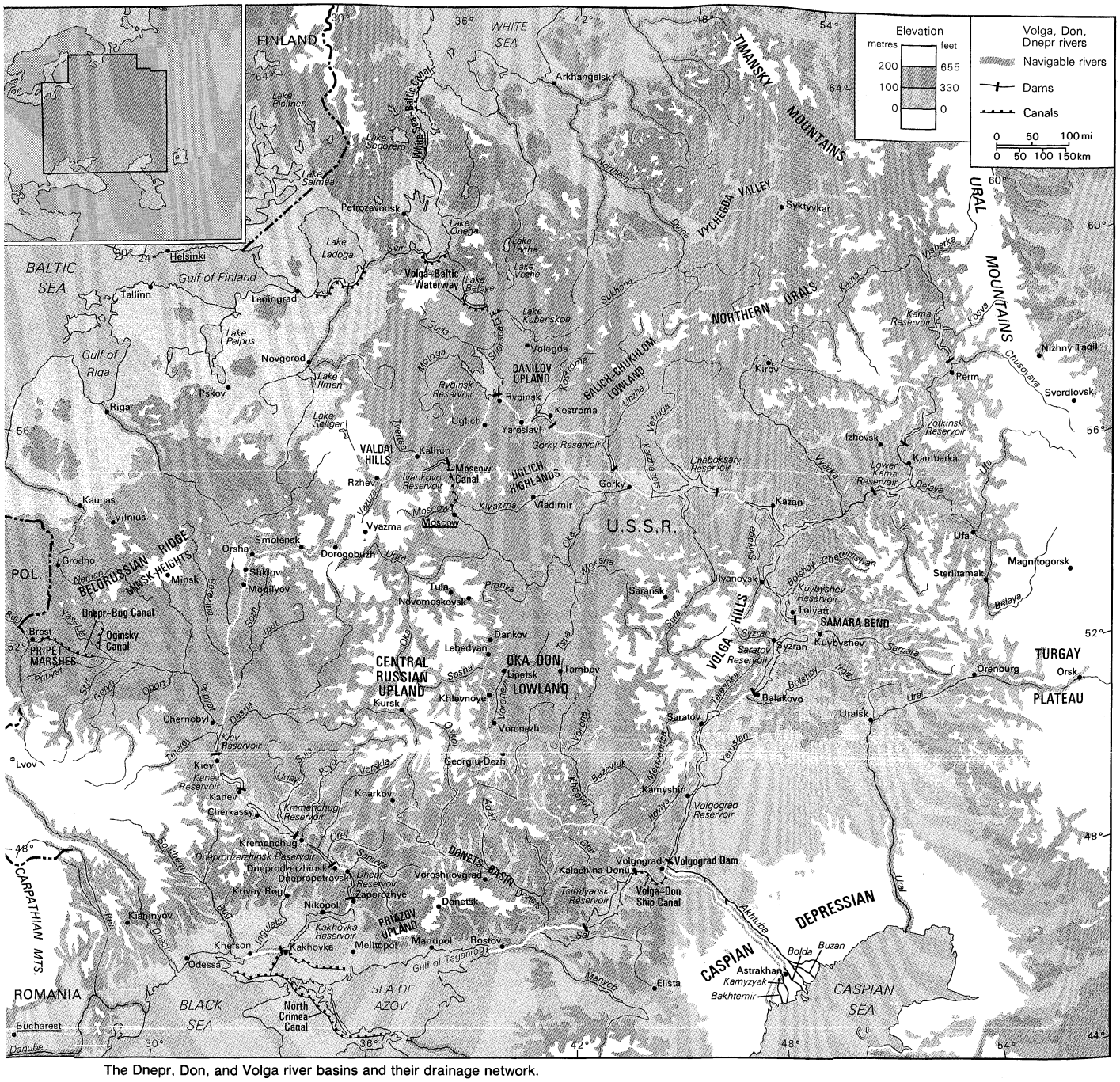
The Dnepr (conventional English: Dnieper; Ukrainian: Dnipro; Belorussian: Dnepro; the Borysthenes of ancient Greek authors) is the third river in length in the European part of the Soviet Union and the fourth longest in Europe, after the Volga, Danube, and Ural rivers. It is 1,367 miles (2,200 kilometres) long and drains an area of about 195,000 square miles (505,000 square kilometres).

The Dnepr rises at an altitude of about 720 feet (220 metres) in a small peat bog on the southern slope of the Valdai Hills, about 150 miles west of Moscow, and flows in a generally southerly direction to the Black Sea. For the first 300 miles, it passes through the Smolensk oblast of the Russian S.F.S.R., first to the south and then to the west; near Orsha it turns south once more and for the next 370 miles flows through the Belorussian S.S.R. Next, it flows through Ukrainian territory: south to Kiev, southeast from Kiev to Dnepropetrovsk, and then south-southwest to the Black Sea.

The Dnepr watershed includes the Volyno-Podolsk Upland, the Belorussian Ridge, the Valdai Hills, the Central Russian Upland, and the Priazov Upland. The centre of the basin consists of broad lowlands. Within the forest area and to some extent within the forest steppe area, the basin is covered with morainic and fluvioglacial deposits; on the steppe it is covered with loess. In some places, where the basin borders upon the basins of the Bug and the Western Dvina rivers, there is a flat swampy area. This facilitated the cutting of connecting water routes from the Dnepr to neighbouring rivers even in ancient times. At the end of the 18th century and the beginning of the 19th, the Dnepr was connected to the Baltic Sea by several canals; the Dnepr–Bug Canal running by way of the Pripyat, Bug, and Vistula rivers; the Oginsky Canal by way of the Pripyat and the Neman; and the Berezina water system by way of the Berezina and the Western Dvina. These canals later became obsolete.

Three sections

Physical features. *Physiography.* The Dnepr is customarily divided into three parts: the upper Dnepr as far as Kiev, the middle Dnepr from Kiev to Zaporozhye (Ukrainian S.S.R.), and the lower Dnepr from Zaporozhye to the mouth. The basin of the upper Dnepr is mainly within a forest area where peat-podzolic soils predominate (replaced in the southern portion of the upper course by podzolized, gray forest soils). The upper Dnepr is characterized by excessive moisture and great swampiness. The river network is well developed in this area, where about four-fifths of the basin's annual runoff forms and the longest tributaries with the greatest runoff (the Berezina, Sozh, Pripyat, Teterev, and Desna) flow. The basin of the middle Dnepr is in a forest steppe area with black earth. Forests stand in the watersheds and along the river valleys. The river network is less dense there, and the rivers carry comparatively less water. The principal tributaries of the middle Dnepr are the Ross, Sula, Psyol, Vorskla, and Samara. The lower Dnepr basin lies within the Black Sea Lowland, in the black-soil steppe area, which has now been completely plowed up. The grassy steppe vegetation



The Dnepr, Don, and Volga river basins and their drainage network.

has been preserved only in the nature reserves and preserves and in old ravines and gullies. Near the Black Sea there is wormwood–fescue vegetation of the semiarid type in chestnut brown soil mixed with saline solonetz and solonchak soils. The lower Dnepr passes through a region of insufficient moisture, where irrigation is employed. The river network there consists for the most part of intermittent streams, the beds of which are ravines that fill with water in the spring and after torrential rains. The largest tributary of this section is the Ingulets.

From its source to Dorogobuzh, Russian S.F.S.R., the Dnepr is a small river flowing past low wooded and, in some places, swampy banks. Downstream the banks rise, and the width of the valley to Orsha varies for the most part from two to six miles, narrowing to less than half a mile in places. Its bed, from 130 to 400 feet wide, is sinuous, with numerous sandbanks. Above Orsha the Dnepr crosses a layer of Devonian limestone, forming the so-called Kobelyaki Rapids, which hamper navigation. From Orsha to Shklov the Dnepr flows between raised, sometimes steep banks overgrown with woods; the left bank becomes lower, whereas the right remains high as far as the confluence with the Sozh River (where the Dnepr enters the Ukrainian S.S.R.). The valley is wide on this stretch, reaching six to nine miles in places. The riverbed from Orsha to Mogilyov is relatively straight; below Mogilyov the Dnepr splits into several channels, producing many islands and sandbanks. The width of the river from Orsha to the confluence with the Sozh ranges from 260 to 1,300 feet, and from the mouth of the Sozh to the mouth of the Pripyat it is from 1,600 to 2,000 feet. The vegetation along the banks of the upper Dnepr consists mainly of wide floodplain meadows, thickets of willows and alders, and old lowland marshes.

Marked asymmetry of the river valley is characteristic of the middle Dnepr. The steep, high right bank (up to 260 feet above the river) forms the escarpment of the Volyno-Podolsk Upland, which stretches along the whole middle course of the river. The low and sloping left bank is formed by broad, ancient terraces. Isolated hills, rising over 300 feet, appear on the low-lying left bank. On the southern portion of the middle Dnepr, the river cuts through the

Ukrainian crystalline masiff and flows for 56 miles in a narrow, almost unterraced valley bounded by high, rocky banks. The Dnepr Rapids, which for centuries prevented continuous navigation, were once located there. The rapids were flooded by the backwaters of the Dnepr hydroelectric power station dam, above Zaporozhye, which raised the level of the river by 130 feet and backed its waters up to Dnepropetrovsk.

The lower river

Below Zaporozhye the Dnepr again passes into a wide valley with a high right bank (130 feet near Nikopol, 260 feet near Kherson). The slopes of the river there are very slight. Before the development of the Kakhovka Reservoir, the waters of which inundated a vast territory, the Dnepr split into a multitude of streams; flat swampy islands, overgrown with floodplain vegetation and reeds, lay among the channels. Today, much of this is hidden under the waters of the reservoir.

Below Kherson the Dnepr forms a delta, the numerous streams of which flow into the Dnepr estuary. Some have been deepened for navigational purposes.

Hydrology. The flow characteristics of the Dnepr have been thoroughly studied. Data on the river's annual runoff date to 1818, while estimates of the maximum discharges—computed from the old high-water marks—extend back more than 250 years. Hundreds of hydrometric stations and posts operate in the Dnepr basin. Under natural conditions the Dnepr had high flows during the spring and fall and low flows during the summer and winter; but dams have altered this regime, so that the river now has pronounced high flows in spring, diminishing flows in summer, and low flows from September to March. Spring snowmelt in the river's upper basin provides the majority of the annual discharge. About 60 percent of the annual runoff occurs from March to May. The period of stable ice on open water in the upper Dnepr sets in at the beginning of December, and in the lower Dnepr at the end of December. Thaw starts at the beginning of April in the upper course and in early March in the lower course. The average annual flow of the river at its mouth is some 59,000 cubic feet (about 1,670 cubic metres) per second; for individual years, the variations in runoff can be considerable. The water of the Dnepr is low in minerals and is soft. In a year the river carries an average of 8.6 million tons of dissolved matter to the sea.

Climate. The climate of the Dnepr basin is, on the whole, temperate and is much milder and damper than that of more eastern regions of the Russian S.F.S.R. located at the same latitude. The continental nature of the climate increases from northwest to southeast. The mean annual air temperature in the upper part of the basin is 41° F (5° C); in the middle (near Kiev), 45° F (7° C); and in the lower reaches of the Dnepr, 50° F (10° C). Winters in the northeast of the basin are long and persistent, whereas in the south they are shorter and milder with frequent thaws; in the north the mean temperature in January is 16° F (−9° C) and in the south 27° F (−3° C). The amount of precipitation decreases from north to south. On the slopes of the Valdai Hills and the Minsk Heights, annual precipitation is about 30–32 inches (760–810 millimetres), while in the lower Dnepr region it is about 18 inches. The mean annual precipitation for the upper Dnepr basin (above Kiev) is about 28 inches. The precipitation average for the entire basin is about 27 inches, with about half falling as rain during the summer and fall.

Temperatures

Plant and animal life. The Dnepr has diverse aquatic flora and fauna. In its upper course the plankton consist mainly of diatom and protococcal algae, rotifers, and *Bosmina.* Blue-green algae come from the mouth of the Pripyat. In its lower course the amount of plankton decreases sharply under the influence of the reservoirs. More than 60 species of fish live in the Dnepr. Commercially important species include pike, roach, chub, ide, rudd, rapfen, tench, barbel, alburnum, golden shiner, goldfish, carp, catfish, burbot, pike perch, perch, and ruff. In the spring the lower Dnepr serves as a habitat for migratory and semimigratory fish (sturgeon, herring, roach, and others). The reservoirs have been stocked artificially with fish of commercial importance, including whitefish, pike perch, golden shiner, and carp.

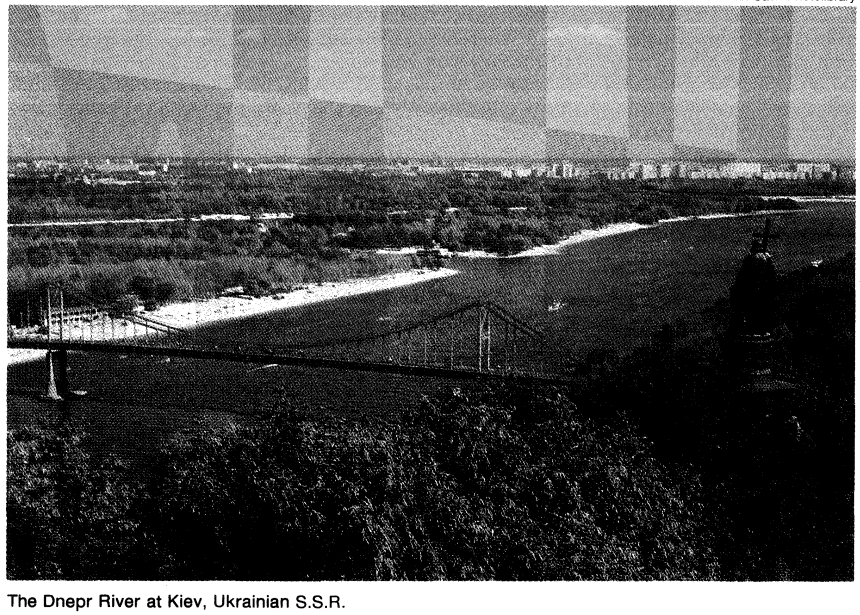
History and economy. The Dnepr basin has been populated since ancient times. It was of central importance in the history of the peoples of eastern Europe, particularly in the founding of the ancient Kievan state. Along this waterway a system of river routes developed in the 4th to 6th century AD a "route from the Varangians to the Greeks," connecting the Black Sea with the Baltic and linking the Slavs with both the Mediterranean and the Baltic peoples. Half of the Dnepr (about 700 miles) borders or passes through territory of the Ukrainian S.S.R., and the river is for the Ukrainians the same kind of national symbol that the Volga River is for the Russians.

The first historical information about the Dnepr is recorded by the Greek historian Herodotus (5th century BC); the river is also mentioned later by the ancient writers Strabo and Pliny the Younger. It was first depicted on a map drawn by Ptolemy in the 2nd century AD. Instrument surveys of the Dnepr were begun early in the 18th century.

River development

Under the Soviets, in line with the general plan for water management, much work has been undertaken for the multipurpose exploitation of the Dnepr's water resources. In 1932, in accordance with the Soviet Union's electrification plan, the river's first hydroelectric power station was completed at Zaporozhye in the region of the rapids. It was the largest power station in Europe until the construc-

J. Allan Cash Photolibrary



The Dnepr River at Kiev, Ukrainian S.S.R.

tion of the huge power stations on the Volga in the 1950s. Completely destroyed by the German army during World War II, the dam was rebuilt in 1947, and its capacity increased. Hydroelectric power stations and reservoirs have also been built on the Dnepr at Kiev (completed 1966), Kanev (1973), Kremenchug (1961), Dneprodzerzhinsk (1965), and Kakhovka (1958). As a result of their construction, many problems have been solved: a continuous deepwater route from the mouth of the Pripyat to the Black Sea has been created; the chronic water shortages in the Donbass and Krivoy Rog industrial regions has been solved; and irrigation of arid lands in the southern Ukraine and the Crimea has been made possible.

Regular navigation on the Dnepr extends as far upstream as Orsha, and, when the water is high, to Dorogobuzh. On the upper Dnepr the required depths are maintained by straightening and by dredging. Below the confluence with the Pripyat, navigable locks make the passage of modern vessels possible. The principal cargoes are coal, ore, mineral building materials, lumber, and grain. The chief ports are Dorogobuzh, Smolensk, Orsha, Mogilyov, Rechitsa, Loyev, Kiev, Cherkassy, Kremenchug, Dnepropetrovsk, Zaporozhye, Nikopol, Kakhovka, and Kherson.

The Krivoy Rog region is supplied with water from the Kakhovka Reservoir by means of the Dnepr–Krivoy Rog Canal. The North Crimea Canal, which was completed in 1971, originates in the reservoir; the canal, 250 miles long, is designed for irrigation of the steppes of the Black Sea Lowland and the northern Crimea and for the creation of a water route from the Dnepr to the Sea of Azov.

Damming the Dnepr and diverting its waters, however, have radically altered its natural hydrology and ecology. Seasonal flow variations have been reduced, upstream access for anadromous fish has been reduced, effluents from cities and industry (as well as from increased agricultural runoff) have caused pollution, and diversion of water for irrigation and evaporation from reservoirs have lowered the annual outflow of the river by some 20 percent. In addition, the wetlands around the river's estuary have been seriously damaged by pollution and reduced discharge.

(A.P.D./P.P.M.)

DON RIVER

One of the great rivers of the European portion of the Soviet Union, the Don has been a vital artery in Russian and Soviet history since the days of Peter I the Great, who initiated a hydrographic survey of its course. Throughout the world the river is associated with images of the turbulent and colourful Don Cossacks—romanticized in a famous series of novels by the 20th century Russian writer Mikhail Sholokhov—and with a series of large-scale engineering projects that have enhanced the waterway's economic importance. The Don rises in the small reservoir of Shat, located in the Central Russian Upland near the city of Novomoskovsk. It flows generally in a southerly direction for 1,162 miles (1,870 kilometres), draining a basin of some 163,000 square miles (422,000 square kilometres), before it enters the Gulf of Taganrog in the Sea of Azov. It is one of the major rivers of the European portion of the Soviet Union, lying between the Volga River to the east and the Dnepr River to the west. In its middle and lower courses, from the confluence with the Chyonaya Kalitva to its mouth, the Don forms an enormous eastward-bulging arc as far as its junction with the Ilovlya. Near the top of the arc, the vast Tsimlyansk Reservoir begins. The Volga–Don Ship Canal stretches from the upper part of the reservoir to the Volga, which at that point is a mere 50 miles distant. From its source in the Tula oblast, the Don crosses the Lipetsk, Voronezh, Volgograd, and Rostov oblasts, through the forest steppe and renowned steppe zones of the Soviet Union. Along the way it collects the waters of numerous tributaries, the most important of which are the Krasivaya Mecha, Sosna, Chyornaya Kalitva, Chir, and Donets (right bank), and the Voronezh, Khopyor, Medveditsa, Ilovlya, Sal, and Manych (left bank). The river winds throughout its course, and the drop along its length is about 620 feet (190 metres).

Important tributaries

Physical features. *Physiography.* In the upper portion of the Don—that is, as far downstream as the southeastward bend—the river flows along the eastern edge of the Central Russian Upland through a generally narrow valley. The right bank is pronounced, reaching heights of 160 feet above the river at the cities of Dankov and Lebedyan, and its limestone and chalk rocks are cut into by ravines and gullies. The left bank borders a flatter floodplain, and the river itself widens intermittently into small lakes; depths range from a few feet in the shoals to 33 feet, with a maximum width of 1,300 feet.

In the middle course, to the beginning of the Tsimlyansk Reservoir, the valley widens to about four miles, and its path is marked by floodplains, more small lakes, and relict channels; the banks, especially the right bank, become steeper, with chalk, limestone, and sandstone predominating. The river narrows to 330–1,300 feet.

The lower course is dominated by the nearly 190 miles of the Tsimlyansk Reservoir, completed in 1953. With an area of some 1,050 square miles and a maximum width of nearly 25 miles, the reservoir has an average depth of about 30 feet. Finally, the lower section of the Don has a valley width of 12–19 miles, with a huge floodplain and a braided river channel as much as 66 feet deep.

The landscape of the upper and middle Don basin is characterized on the right bank by undulating plains cut into by jagged gorges and on the left bank by the smoother, pond-dotted topography of the Oka–Don Lowland. Farther downriver the vast open landscapes of the steppes predominate. Rich, black chernozem soils fill almost the entire basin, though there are patches of gray forest soil in the north, where forests cover up to 12 percent of the area.

Hydrology. The long-term fluctuations in the water level of the Don reach about 40 feet in the upper course, 25 feet in the middle course, and 20 feet in the lower course. The highest levels are in the spring, the lowest in autumn and winter. At the mouth of the Don strong winds from the sea cause increases in the water level (wind surges). The average rate of discharge at the mouth of the Don is about 31,800 cubic feet (900 cubic metres) per second, but the river experiences great variation in its flow during the year. For example, at the city of Georgy-Dezh, in the river's upper course, the average flow is about 8,900 cubic feet per second, but flows range from 1,500 to approximately 395,000 cubic feet per second. There are corresponding variations as the annual flow increases downstream. At the city of Kalach-na-Donu about 65 percent of the annual flow occurs during April and May, compared with about 7 percent in March before the snowmelt begins. Below the Tsimlyansk Reservoir the flow has been partially regulated. At Nikolayevskaya, for example, 34 percent of the annual volume occurs in spring, 33 percent in summer, 22 percent in autumn, and 11 percent in winter.

Fluctuating water levels

The northern portion of the Don begins to freeze by about mid-November and is clear of ice by mid-April. In the lower course the river is frozen from the end of November to the end of March at Kalach-na-Donu and from mid-December to the beginning of April at Rostov-na-Donu.

Climate. The climate of the basin is moderately continental, with average January temperatures ranging from 12° F to 18° F (−11° C to −8° C), while July readings reach 66° F to 72° F (19° C to 22° C). Annual precipitation diminishes from 23 inches (584 millimetres) in the north to 14–15 inches in the south.

History and economy. Archaeological evidence of early settlement of the Don River basin dates from the Upper Paleolithic (40,000–13,000 years ago). At the beginning of the 2nd century BC, tribes of herdsmen occupied the valley of the Don and developed livestock raising and crop agriculture there. The Tatars conquered the region during the first half of the 13th century AD. The Russian state, expanding southward from the Grand Principality of Moscow, incorporated the Don River basin between the middle of the 15th and 16th centuries. The famed Don Cossacks established themselves in independent military settlements along the middle and lower Don by the 16th century but subsequently came under tsarist control.

Since the early 1950s the Don has undergone intensive

Quays along the Don River at Rostov-na-Donu, Russian S.F.S.R.
M. Koene—H. Armstrong Roberts

Construction of reservoirs

economic development. The key to this was the creation of the huge Tsimlyansk Reservoir along its lower course. The project included a hydroelectric station, a fish elevator, two navigation locks, an irrigation canal, a 1,580-foot concrete dam, and an eight-mile earthen dam. By 1975 an additional 116 reservoirs, with volumes exceeding 35 million cubic feet each, existed in the basin.

The Tsimlyansk Reservoir contributed to a rapid expansion of irrigation in the Don River basin, which grew from about 124,000 acres (50,000 hectares) in 1950 to nearly 2.5 million acres by 1980. In the upper basin an extensive network of ponds aids irrigation; these ponds are also used for raising fish.

The significance of the Don as a navigable waterway greatly increased with the construction of the Volga–Don Ship Canal. The river itself is navigable from the mouth to the city of Georgy-Dezh (a distance of 842 miles) and in the spring for another 150 miles upstream. Navigation in the lower course has been facilitated greatly by the Tsimlyansk project. Navigation at the mouth of the Don is occasionally hindered by the declines in water level induced by strong, persistent offshore winds, while dredging operations are necessary to maintain and improve navigation in the upper reaches. The largest ports are Kalach-na-Donu, Tsimlyansk, and Rostov-na-Donu.

The development of the Don has provided substantial economic benefits to the riverine populations as well as to the nation, but these alterations have reduced substantially the amount of water discharged at the river's mouth. This decrease—estimated in 1975 to be 20 percent of the 1950 level and still rising—has come chiefly from water diversion for irrigation and through evaporation from the artificial reservoirs; and, as a result of it, the salinity of the Sea of Azov has risen considerably, diminishing the sea's biological productivity and lowering fish catches.

(A.M.Ga./P.P.M.)

VOLGA RIVER

The significance of the river

Europe's longest river, the Volga (ancient Ra, medieval Itil or Etil) is the historic cradle of the Russian state. Its basin, sprawling across a third of what is now the European part of the Soviet Union, contains a quarter of the entire Soviet population. The Volga's immense economic, cultural, and historic importance—along with the sheer size of the river and its basin—ranks it among the world's great rivers. Rising in the Valdai Hills northwest of Moscow, the Volga discharges into the Caspian Sea, some 2,193 miles (3,530 kilometres) to the south. It drops slowly and majestically from its source 748 feet (228 metres) above sea level to its mouth 92 feet below sea level. In the process the Volga receives the water of some 200 tributaries, the majority of which join the river on its left bank. Its river system, comprising 151,000 rivers and permanent and intermittent streams, has a total length of 357,000 miles.

Physical features. The river basin—533,000 square miles (1,380,000 square kilometres) in area—drains one-third of the Soviet Union's European territory, stretching from the Valdai Hills and Central Russian Upland in the west to the Ural Mountains in the east, and narrowing sharply at Saratov in the south. From Kamyshin the river flows to its mouth uninterrupted by tributaries for some 400 miles.

Four geographic zones lie within the Volga basin: the dense, marshy forest, which extends from the river's upper reaches to Gorky and Kazan; the forest steppe extending from there to Kuybyshev and Saratov; the steppe from there to Volgograd; and semidesert lowlands southeast to the Caspian Sea.

Physiography. The course of the Volga is divided into three parts: the upper Volga (from its source to the confluence of the Oka); the middle Volga (from the confluence of the Oka to that of the Kama); and the lower Volga (from the confluence of the Kama to the mouth of the Volga itself). The Volga is a small stream in its upper course through the Valdai Hills, becoming a true river only after the entrance of several of its tributaries. It then passes through a chain of small lakes, receives the waters of the Selizharovka River, and then flows southeast through a terraced trench. Past the town of Rzhev, the Volga turns northeastward, is swelled by the inflow of the Vazuza and Tvertsa rivers at Kalinin, and then continues to flow northeastward through the Rybinsk Reservoir, into which other rivers, such as the Mologa and the Sheksna, flow. From the reservoir the river proceeds southeastward through a narrow, tree-lined valley between the Uglich Highlands to the south and the Danilov Upland and the Galich-Chukhlom Lowland to the north, continuing its course along the Unzha and the Balakhna lowlands to Gorky. (Within this stretch the Kostroma, Unzha, and Oka rivers enter the Volga.) On its east-southeastward course from the confluence of the Oka to Kazan, the Volga doubles in size, receiving waters from the Sura and Sviyaga on its right bank and the Kerzhenets and Vetluga on its left. At Kazan the river turns south into the Kuybyshev Reservoir, where it is joined from the left by its major tributary, the Kama. From this point the Volga becomes a mighty river, which, save for a sharp loop at the Samara Bend, flows southwestward along the foot of the Volga Hills in the direction of Volgograd. (Between the Samara Bend and Volgograd it receives only the relatively small left-bank tributaries of the Samara, Bolshoy Irgiz, and Yeruslan.) Above Volgograd the Volga's main distributary, the Akhtuba, branches southeastward to the Caspian Sea, running parallel to the main course of the

river, which also turns southeast. A floodplain, characterized by numerous interconnecting channels and old cutoff courses and loops, lies between the Volga and the Akhtuba. Above Astrakhan a second distributary, the Buzan, marks the beginning of the Volga delta, which, with an area of more than 7,330 square miles, is the largest in the Soviet Union. Other main branches of the Volga delta are the Bakhtemir, Kamyzyak, Staraya (Old) Volga, and Bolda.

Hydrology. The Volga is fed by snow (which accounts for 60 percent of its annual discharge), underground water (30 percent), and rainwater (10 percent). The natural, untamed regime of the river was characterized by high spring floods (*polovodye*). Before it was regulated by reservoirs, annual fluctuations in level ranged from 23 to 36 feet on the upper Volga; from 39 to 46 feet on the middle Volga; and from 10 to 49 feet on the lower Volga. At Kalinin the average annual rate of river flow is about 6,400 cubic feet (180 cubic metres) per second, at Yaroslavl 39,000 cubic feet per second, at Kuybyshev 272,500 cubic feet per second, and at the river's mouth 284,500 cubic feet per second. Below Volgograd the river loses about 2 percent of its waters in evaporation. More than 90 percent of annual runoff occurs above the confluence of the Kama. Increased flow below the Kama is only about 7 percent.

Former fluctuations in water level

Climate. The climate of the Volga basin changes significantly from north to south. From its source to the Kama confluence, it lies within a temperate climatic zone characterized by a cold, snowy winter and a warm, rather humid summer. From the Kama to below the Volga Hills, hot, dry summers and cold winters with little snow prevail. Temperatures and evaporation increase, and precipitation decreases, toward the south and east: the average January temperatures in the river's upper reaches range from 19° F (−7° C) to 6° F (−14° C) and those of July from 62° F (17° C) to 68° F (20° C), while on its lower reaches at Astrakhan corresponding temperatures are 19° F (−7° C) and 77° F (25° C). Annual rainfall ranges from 25 inches (635 millimetres) on the northwest to 12 inches on the southeast. Evaporation from land ranges from 20 inches in the northwest to eight inches in the southeast. The upper and middle courses of the Volga begin to freeze at the end of November, the lower reaches in December. At Astrakhan the ice breaks up in mid-March, at Kamyshin at the beginning of April, and everywhere else in mid-April. The Volga is generally free of ice for about 200 days each year and near Astrakhan for about 260 days. As great masses of water accumulated within the reservoirs constructed during the Soviet period, however, the temperature regime of the Volga was so changed that the duration of ice increased on the headwaters of the reservoirs and decreased on the stretches below the dams.

The changing river regime

The economy. *Dams and reservoirs.* A string of huge dams and reservoirs now line the Volga and its major tributary, the Kama River, converting them from free-flowing rivers to chains of man-made lakes. All of the reservoir complexes include hydroelectric power stations and navigation locks. The uppermost complex on the Volga, the Ivankovo , with a reservoir covering 126 square miles, was completed in 1937, and the next complex, at Uglich (96 square miles), was put into operation in 1939. The Rybinsk Reservoir, completed in 1941 and encompassing an area of about 1,750 square miles, was the first of the large reservoir projects. Following World War II, work continued below Rybinsk. The Gorky and Kuybyshev reservoirs were both completed in 1957, and the Cheboksary Reservoir, located between them, became operational in 1980. The huge Kuybyshev Reservoir, with an area of some 2,300 square miles, is the largest element of the Volga reservoir system; it not only impounds the waters of the Volga but also backs water up the Kama for some 375 miles. The Saratov and Volgograd reservoirs (completed in 1968 and 1962, respectively) are the last such bodies on the Volga itself. The chain on the Kama consists of three reservoirs, the newest of which—the Lower Kama Reservoir—became operational in 1979. There are a total of eight hydroelectric stations on the Volga and three on the Kama, which combined have an installed generating capacity of some 11 million kilowatts of power.

Navigation. The Volga, navigable for some 2,000 miles, and its more than 70 navigable tributaries carry more than half of all Soviet inland freight and nearly half of all the passengers who use Soviet inland waterways. Construction materials and raw materials account for about 80 percent of the total freight; other cargoes include petroleum and petroleum products, coal, foodstuffs, salt, tractors and agricultural machinery, automobiles, chemical apparatus, and fertilizers. The major ports on the Volga are Kalinin, Rybinsk, Yaroslavl, Gorky, Kazan, Ulyanovsk, Kuybyshev, Saratov, Kamyshin, Volgograd, and Astrakhan.

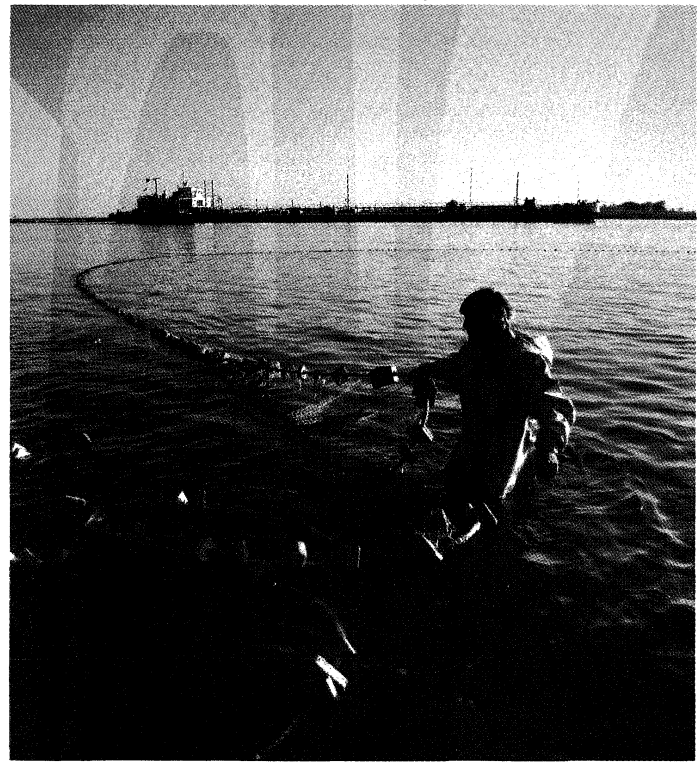
Freight patterns

The Volga is joined to the Baltic Sea by the Volga–Baltic Waterway, which, in turn, is joined to the White Sea (via Lake Onega) by the White Sea–Baltic Canal; to the Moscow River, and hence to Moscow, by the Moscow Canal; and to the Sea of Azov by the Volga–Don Ship Canal. The river has thus become integrated in a vital way with virtually the entire waterway system of the European Soviet Union.

Environmental changes. Although the extensive development of the Volga has made a major contribution to the Soviet economy, it also has had adverse ecological consequences. The system of dams and reservoirs has blocked or severely curtailed access for such anadromous species as the beluga sturgeon (famous for the caviar made from its roe) and whitefish (*belorybitsa*), which live in the Caspian Sea but spawn in the Volga and other inflowing rivers, and it has fundamentally altered the habitat of the nearly 70 species of fish native to the river. These changes—along with pollution by industrial and municipal effluents and by agricultural runoff—have led to deterioration of the major Volga fisheries. Water loss by impoundment and evaporation and by diversion (chiefly for irrigation) have diminished discharge at the mouth of the Volga compared with natural conditions, and this has contributed to an almost steady decline in the level of the Caspian Sea since 1930. Intensive efforts to alleviate these man-made influences, however, have been under way for a number of years. For example, some three-fifths of the Caspian sturgeon are now bred artificially rather than in their natural spawning grounds.

Study and exploration. The Volga was known to the Alexandrian geographer Ptolemy (2nd century AD), to the Slavs, and to the Arab geographers of the 10th and 11th

Jonathan Wright—Bruce Coleman Inc.



Fishing for beluga sturgeon in the Volga River, Volgograd, Russian S.F.S.R.

centuries. Information on it is contained in the *Kniga bolshomu chertyozhu* (1627; "Book of the Great Chart") and in a hydrographic description of 1636. Its flow was first measured below Kamyshin by the Englishman John Perry in 1700. Two pioneer Russian navigators, Makeyev and Gavril Andreyevich Sarychev, surveyed the stretch between Tver (now Kalinin) and Nizhny Novgorod (now Gorky) in 1782–83; in 1809–17 and 1829 the Maritime Bureau surveyed the delta and measured its depth; and from 1875 to 1894 the river was investigated from the Rybinsk to the Volga mouth. Investigations of the upper Volga were made from 1896 to 1901, and in 1894 the upper reaches of the Volga, Oka, Syzran, and other rivers were also examined. Many institutes have carried out hydrographic and hydrometric research during the Soviet period: by the late 20th century there were more than 500 points at which the water levels of the Volga had been observed. (P.S.K./P.P.M)

BIBLIOGRAPHY. *General:* Sources that provide brief but comprehensive information on European states include *Western Europe 1989: A Political and Economic Survey* (1988), from Europa Publications; and two surveys from "The World Today Series": WAYNE C. THOMPSON and MARK H. MULLIN, *Western Europe 1988,* 7th annual ed. (1988); and M. WESLEY SHOEMAKER, *The Soviet Union and Eastern Europe 1988,* 19th annual ed. (1988). RICHARD MAYNE (ed.), *Western Europe,* rev. ed. (1987); and GEORGE SCHÖPFLIN (ed.), *The Soviet Union and Eastern Europe,* rev. ed. (1986), both from the series "Handbooks to the Modern World," are more detailed analyses. DENYS HAY, *Europe: The Emergence of an Idea,* rev. ed. (1968), is a work of historical geography that explores the concept "Europe." Other historical works include GORDON EAST, *An Historical Geography of Europe,* 5th ed. (1966); and NORMAN J.G. POUNDS, *An Historical Geography of Europe, 450 B.C.–A.D. 1330* (1973), *An Historical Geography of Europe, 1500–1840* (1979), and *An Historical Geography of Europe, 1800–1914* (1985). Annuals include *The Statesman's Year-Book* and UNITED NATIONS, *Statistical Yearbook.* (T.M.P.)

Physical and human geography: (*Geologic history*): A survey of the geology of the continent is offered in DEREK V. AGER, *The Geology of Europe: A Regional Approach* (1980). ROLAND BRINKMANN, *Geologic Evolution of Europe,* 2nd rev. ed. (1969; originally published in German, 8th ed., 1959), is an introductory summary. DEREK V. AGER and M. BROOKS (eds.), *Europe from Crust to Core* (1977), collects papers on geologic events, from oldest to youngest, presented at a meeting of European geologic societies. M.G. RUTTEN, *The Geology of Western Europe* (1969), provides a general geologic background of part of the continent. Basic geologic elements are discussed in two articles published in *Geologie en mijnbouw,* vol. 57, no. 4 (1978): PETER A. ZIEGLER, "North-Western Europe: Tectonics and Basin Development," pp. 589–626; and H.J. ZWART and U.F. DORNSIEPEN, "The Tectonic Framework of Central and Western Europe," pp. 627–654. D.V. NALIVKIN, *Geology of the U.S.S.R.* (1973; originally published in Russian, 1962), includes substantial coverage of the European part of the country. Beautiful colour maps illustrating the evolution of Europe are found in PETER A. ZIEGLER, *Geological Atlas of Western and Central Europe* (1982). (B.F.W.)

(*The land*): General discussions of such topics as climate, topography, relief, vegetation zones, and animal distribution are found in GEORGE W. HOFFMAN (ed.), *A Geography of Europe,* 5th ed. (1983); TERRY G. JORDAN, *The European Culture Area,* 2nd ed. (1988); MARGARET REID SHACKLETON, *Europe, a Regional Geography,* 7th enlarged ed., rev. by GORDON EAST (1969); F.J. MONKHOUSE, *A Regional Geography of Western Europe,* 4th ed. (1974); and E.C. MARCHANT (comp.), *The Countries of Europe as Seen by Their Geographers* (1970). See also "Europe (Excluding Russia)," pp. 297–388 in W.G. KENDREW, *The Climate of the Continents,* 5th ed. (1961).

Works that focus on the geography of specific regions of Europe include BRIAN S. JOHN, *Scandinavia* (1984); ROY E.H. MELLOR, *The Two Germanies* (1978); D.S. WALKER, *The Mediterranean Lands,* 3rd ed. (1965); J.M. HOUSTON, *The Western Mediterranean World* (1964); NORMAN J.G. POUNDS, *Eastern Europe* (1969); DEAN S. RUGG, *Eastern Europe* (1985); PAUL E. LYDOLPH, *Geography of the U.S.S.R.* (1979); and LESLIE SYMONS *et al., The Soviet Union, a Systematic Geography* (1983).

(*People*): Historical development of anthropological and ethnological characteristics is outlined in TIMOTHY CHAMPION *et al., Prehistoric Europe* (1984); CARLETON STEVENS COON, *The Races of Europe* (1939, reprinted 1972); MICHAEL W. FLINN, *The European Demographic System, 1500–1820* (1981); and JOHN GEIPEL, *The Europeans: An Ethnohistorical Survey* (1969). BRIAN W. ILBERY, *Western Europe: A Systematic Human Geography,* 2nd ed. (1986), is a concise overview. Population trends of Europe in relation to those of the other continents are discussed in J. BEAUJEU-GARNIER, *Geography of Population,* 2nd ed. (1978; originally published in French, 2 vol., 1956–58). For statistical information, UNITED NATIONS, *Demographic Yearbook,* is useful. The growing minority nationalist movements are examined in CHARLES R. FOSTER (ed.), *Nations Without a State: Ethnic Minorities in Western Europe* (1980); HUGH SETON-WATSON, *Nations and States: An Enquiry into the Origins of Nations and the Politics of Nationalism* (1977); LOUIS L. SNYDER, *Global Mini-Nationalisms: Autonomy or Independence* (1982); GEORGE KLEIN and MILAN J. REBAN (eds.), *The Politics of Ethnicity in Eastern Europe* (1981); and STEPHEN CASTLES, *Here for Good: Western Europe's New Ethnic Minorities* (1984), which focuses on the problems of foreign labour forces. A broad range of other topics is treated in such special studies as STANLEY HOFFMANN and PASCHALIS KITROMILIDES (eds.), *Culture and Society in Contemporary Europe* (1981); JAN F. TRISKA and CHARLES GATI (eds.), *Blue-Collar Workers in Eastern Europe* (1981); S.H. FRANKLIN, *The European Peasantry: The Final Phase* (1969); DAVID LANE, *The End of Social Inequality?: Class, Status, and Power Under State Socialism* (1982); RICHARD T. DE GEORGE and JAMES P. SCANLAN (eds.), *Marxism and Religion in Eastern Europe* (1975); and VERNON MALLINSON, *The Western European Idea in Education* (1980).

(*Economy*): An introduction to European economic history is useful for understanding the modern European economy. The ongoing multivolume series "Cambridge Economic History of Europe," begun in the 1960s under the general editorship of M.M. POSTAN, provides comprehensive surveys. Important historical periods are explored in HARRY A. MISKIMIN, *The Economy of Early Renaissance Europe, 1300–1460* (1975), and *The Economy of Later Renaissance Europe, 1460–1600* (1977); CARLO M. CIPOLLA, *Before the Industrial Revolution: European Society and Economy, 1000–1700,* 2nd ed. (1980; originally published in Italian, 1974); A.G. KENWOOD and A.L. LOUGHEED, *The Growth of the International Economy, 1820–1980* (1983); and M.C. KASER (ed.), *The Economic History of Eastern Europe, 1919–1975,* 3 vol. (1986–87).

General analyses of the economic character of Europe include HUGH CLOUT, *Regional Development in Western Europe,* 3rd ed. (1987); WALTER LAQUEUR, *A Continent Astray: Europe, 1970–1978* (1979); ANDREA BOLTHO (ed.), *The European Economy* (1982); ANDREW J. PIERRE (ed.), *Unemployment and Growth in the Western Economies* (1984); JOZEF M. VAN BRABANT, *Socialist Economic Integration: Aspects of Contemporary Economic Problems in Eastern Europe* (1980); ALAN H. SMITH, *The Planned Economies of Eastern Europe* (1983); and PAUL STONHAM, *Major Stock Markets of Europe* (1982). For current information on a diversity of economic topics, UNITED NATIONS, *Economic Survey of Europe* (annual), is useful. European agriculture is discussed in MICHAEL TRACY, *Government and Agriculture in Western Europe, 1880–1988,* 3rd ed. (1989); RUTH ELLESON, *Performance and Structure of Agriculture in Western Europe* (1983); KARL-EUGEN WÄDEKIN, *Agrarian Policies in Communist Europe* (1982); and ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Prospects for Agricultural Production and Trade in Eastern Europe,* 2 vol. (1981–82). Industry, technology, and energy are the special focus of GEOFFREY SHEPHERD, FRANÇOIS DUCHÊNE, and CHRISTOPHER SAUNDERS (eds.), *Europe's Industries* (1983); and GEORGE W. HOFFMAN, *The European Energy Challenge* (1985).

WILLIAM ASHWORTH, *A Short History of the International Economy Since 1850,* 4th ed. (1987), provides an introduction to the idea of economic cooperation; and cooperation is further explored in JULIET LODGE (ed.), *Institutions and Policies of the European Community* (1983); PETER LUDLOW, *The Making of the European Monetary System* (1982); DENNIS SWANN, *Competition and Industrial Policy in the European Community* (1983); and VALERIE J. ASSETTO, *The Soviet Bloc in the IMF and the IBRD* (1988). (T.M.P.)

Special geographic features: Literature on the geographic features of Europe is often sketchy or technical. For general overviews of the features discussed in the article, the reader is advised to turn to books cited above in the *Physical and human geography* section of the bibliography.

(*Landforms*): General descriptive studies of the Alps include PAUL VEYRET and GERMAINE VEYRET, *Au Coeur de l'Europe, les Alpes* (1967), and PAUL VEYRET, *Les Alpes* (1972); GÜNTER GLAUERT, *Die Alpen, eine Einführung in die Landeskunde* (1975); and *The Alpes* (1984), an illustrated multilanguage work published under the auspices of the 25th International Geographical Congress. Works on physical geography include LÉON W. COLLET, *The Structure of the Alps,* 2nd ed. (1935, reprinted 1974), which sets forth the theory of the nappes; ERNST KRAUS, *Die Baugeschichte der Alpen,* 2 vol. (1951), which provides a geologic synthesis; and ALBRECHT PENCK and EDUARD BRÜCKNER, *Die Alpen im Eiszeitalter,* 3 vol. (1901–09), which traces

the history of glaciation. For human geography, see PIERRE GABERT, *Les Alpes et les états alpins* (1965); MICHEL CÉPÈDE and E.S. ABENSOUR, *Rural Problems in the Alpine Region, an International Study* (1961); and PIER PAOLO VIAZZO, *Upland Communities: Environment, Population, and Social Structure in the Alps Since the Sixteenth Century* (1989). PIERRE GEORGE and JEAN TRICART, *L'Europe centrale* (1954), includes information on economic development of the region in the first half of the 20th century; and LOUIS CHABERT, *Les Grandes Alpes industrielles de Savoie: évolution économique et humaine* (1978), is a regional socioeconomic analysis. Other regional economic studies include PAUL VEYRET and GERMAINE VEYRET, *Atlas et géographie des Alpes françaises* (1979); AUBREY DIEM (ed.), *The Mont Blanc-Pennine Region* (1984); ERNST A. BRUGGER *et al.* (eds.), *The Transformation of Swiss Mountain Regions* (1984), a detailed survey; AUBREY DIEM, *Switzerland, Land, People, Economy* (1986); BERNARD JANIN, *Une Région alpine originale, la Val d'Aoste,* 2nd rev. ed. (1976), a descriptive work with an economic focus; and ELISABETH LICHTENBERGER, *The Eastern Alps* (1975), a brief description in the series "Problem Regions of Europe"; and MARY L. BARKER, "Traditional Landscape and Mass Tourism in the Alps," *Geographical Review* 72(4):395–415 (October 1982). Specific features of the Alpine economy, especially agriculture, are addressed in JOHN FRÖDIN, *Zentraleuropas alpwirtschaft,* 2 vol. (1940–41); and H. AULITZKY, *Endangered Alpine Regions and Disaster Prevention Measures* (1974). The historical character of the region is explored in PAUL GUICHONNET (ed.), *Histoire et civilisations des Alpes,* 2 vol. (1980); and LUDWIG PAULI, *The Alps: Archaeology and Early History* (1984; originally published in German, 1980). (A.Di.)

Literature on the Apennines includes D. POSTPISCHL (ed.), *Catalogo dei terremoti italiani dall'anno 1000 al 1980* (1985), a scientific catalog of earthquakes, with an extended abstract in English that provides information on geologic characteristics of the range; CALVINO GASPARINI, ENRICO GIORGETTI, and MAURIZIO PAROTTO, *Il terremoto in Italia: cause, salvaguardia, interventi* (1984), a study of the seismic hazards in the region and of protective measures against them; SANDRO PIGNATTI, *Flora d'Italia,* 3 vol. (1982), a discussion of the major plants of the area; J.M. SCOTT, *A Walk Along the Apennines* (1973), which offers a description of views and localities; and ROLAND SARTI, *Long Live the Strong: A History of Rural Society in the Apennine Mountains* (1985). Works that contain detailed geologic information on the Apennines include L. OGNIBEN, M. PAROTTO, and A. PRATURLON (eds.), *Structural Model of Italy: Maps and Explanatory Notes* (1975), and *Cento anni di geologia italiana* (1981), a centennial publication of the Italian Geological Society. (Ma.P.)

Thorough, though sometimes brief, treatments of the Carpathians are found in regional geographies such as EMMANUEL DE MARTONNE, *Europe centrale,* 2 vol. (1930–31); MÁRTON PÉCSI and BÉLA SÁRFALVI, *The Geography of Hungary* (1964); TIBERIU MORARIU, VASILE CUCU, and ION VELCEA, *The Geography of Romania,* 2nd ed. (1969); JAROMÍR DEMEK *et al., Geography of Czechoslovakia,* trans. from Czech (1971); and IRENA KOSTROWICKA and JERZY KOSTROWICKI, *Poland: Landscape and Architecture* (1980; originally published in Polish, 1969). Specifically on the Carpathians, G.Z. FÖLDVARY, *Geology of the Carpathian Region* (1988), is informative and detailed, though technical. The Carpathian region is one of the three mountain regions discussed in P. SKALNIK, "Uneven and Combined Development in European Mountain Communities," pp. 123–154 in DAVID C. PITT (ed.), *Society and Environment, the Crisis in the Mountains* (1978). (J.A.K.)

Research articles on the Pyrenees appear in such journals as *Pyrénées* (quarterly), published by the Musée Pyrénéen du Château-Fort de Lourdes; *Revue géographique des Pyrénées et du Sud-Ouest* (quarterly); *Annales du Midi* (five times a year); and *Pirineos: publicación de la Estación de Estudios Pirenaicos* (annual). General surveys include HENRY MYHILL, *The Spanish Pyrenees* (1966); FRANÇOIS TAILLEFER (ed.), *Les Pyrénées: de la montagne à l'homme* (1974); GEORGES VIERS, *Les Pyrénées,* 3rd ed. (1973); and CLAUDE DENDALETCHE, *Pyrénées* (1982). ROGER HIGHAM, *Road to the Pyrenees* (1971); and J.M. SCOTT, *From Sea to Ocean: Walking Along the Pyrenees* (1969), are descriptive works based on travel experiences. PAUL G. BAHN, *Pyrenean Prehistory: A Palaeoeconomic Survey of the French Sites* (1983); and DANIEL ALEXANDER GÓMEZ-IBÁÑEZ, *The Western Pyrenees: Differential Evolution of the French and Spanish Borderland* (1975), are historical geographies. For human geography, see MICHEL CHEVALIER, *La Vie humaine dans les Pyrénées ariégeoises* (1956); LLUÍS SOLÉ I SABARÍS, *Los Pirineos: el medio y el hombre* (1951); and NEIL LANDS, *History, People, and Places in the French Pyrenees* (1980). (F.O.)

Sources on the Urals in Western languages are scarce. I.V. KOMAR and A.G. CHIKISHEV (eds.), *Урал и Приуралье* (1968), is a comprehensive survey of relief, geology, climate, drainage, soils, flora, and fauna of the region, with data on natural resources, economic development, and preservation of the environment. A.A. MAKUNINA, *Ландшафты Урала* (1974), deals specifically with the geomorphology of the region. N.P. ARKHIPOVA and E.V. YASTREBOV, *Как были открыты Уральские горы* (1971), is the history of the discovery and development of the Ural mountain region. B. RYABININ, *Across the Urals,* trans. from Russian (1973), is a descriptive work based on travels in the area. M.T. IOVCHUK and L.N. KOGAN (eds.), *The Cultural Life of the Soviet Worker: A Sociological Study* (1975), offers a glimpse of working-class life in this highly developed industrial region. (Y.V.Y.)

(*Western European drainage systems*): On the Rhine, see WILLIAM GRAVES, "The Rhine: Europe's River of Legend," *National Geographic* 131(4):449–499 (April 1967), based on a voyage aboard a Rhine tanker from Rotterdam to Karlsruhe. GORONWY REES, *The Rhine* (1967), is a longer description, which follows the Rhine from its source to its mouth and includes historical, political, cultural, and economic information. ROYAL INSTITUTE OF INTERNATIONAL AFFAIRS, *Regional Management of the Rhine* (1975), is a collection of scholarly but readable papers on the effects of human activity on the ecology of the river, with analyses of transport, navigation, flood control, pollution, generation of electricity, regional planning, and recreational use. H.J. MACKINDER, *The Rhine, Its Valley and History* (1908), is a classic study by one of the founders of modern academic geography, still worth reading. E.M. YATES, "The Development of the Rhine," *Transactions, Institute of British Geographers,* publication no. 32, pp. 65–81 (1963), examines the physical evolution of the Rhine and its valley from the Oligocene to the end of the Ice Age. ROY E.H. MELLOR, *The Rhine: A Study in the Geography of Water Transport* (1983), surveys the history of navigation on the river. Fuller systematic treatments, which include discussions of the history of economic activity of the region, population dynamics, and political and cultural developments, are ÉTIENNE JUILLARD, *L'Europe rhénane* (1968); and JEAN DOLLFUS, *L'Homme et le Rhin* (1960). (K.A.Si.)

Much of what has been written on the Rhône is included in general and regional geographies of Switzerland, France, and western Europe, such as AUBREY DIEM, *Western Europe, a Geographical Analysis* (1979). DANIEL FAUCHER, *L'Homme et le Rhône* (1968), provides a historical survey of water resources development and economic conditions. To supplement it, see the earlier exhaustive works by a hydrologist of world reputation, MAURICE PARDE, *Le Régime du Rhône: étude hydrologique* (1925), continued in his *Quelques Nouveautées sur le régime du Rhône* (1942). A short account that focuses on economic conditions, from the series "Problem Regions of Europe," is IAN B. THOMPSON, *The Lower Rhône and Marseille* (1975). (A.Di.)

The earliest scientific work on the Seine is EUGÈNE BELGRAND, *La Seine, études hydrologiques: régime de la pluie, des sources, des eaux courantes* (1872), with an accompanying *Atlas* (1873), which is still valuable despite its age. Development of navigation on the river is surveyed in AIMÉ V. PERPILLOU, "Un Exemple de canalisation de rivière: la Seine," pp. 37–49 in the author's *Géographie de la circulation: conditions générales de la navigation intérieure* (1950). JACQUES GRAS, *Le Bassin de Paris méridional* (1963), examines the morphology of the Paris and Loire basins, as well as of the Loing valley and part of the Yonne basin. Useful information on the Seine basin is found in *Les Bassins de la Seine et des cours d'eau Normands* (1975), published by Agence Financière de Bassin Seine-Normandie. Available English-language sources include such travel books as ANTHONY GLYN, *The Seine* (1966); and WILLIAM DAVENPORT, *The Seine: From Its Source, to Paris, to the Sea* (1968). EVELYN BERNETTE ACKERMAN, *Village on the Seine: Tradition and Change in Bonnières, 1815–1914* (1978), is a scholarly examination of history and socioeconomic conditions as influenced by the river. (M.Da.)

(*Central European drainage systems*): Useful works include JOSEF BREU, *Atlas of the Danubian Countries,* 11 issues in 2 vol. (1970–89), a comprehensive, multilingual source on the Danube region's geography; and ANTON SIKORA, LUDOVÍT ÜRGE, and DOMOKOS MIKLÓS, *Danube* (1988). Much of the literature in English on the Danube itself consists of descriptive works based on travel experiences, such as PATRICK LEIGH FERMOR, *Between the Woods and the Water: On Foot to Constantinople from the Hook of Holland: The Middle Danube to the Iron Gates* (1986); and CLAUDIO MAGRIS, *Danube* (1989; originally published in Italian, 1986). Navigation of the river and its influence on the economic development of the region are surveyed in J.P. CHAMBERLAIN, *The Regime of the International Rivers: Danube and Rhine* (1923, reprinted 1968); and STEPHEN GOROVE, *Law and Politics of the Danube* (1964), discusses the regulations of navigation and the river's international importance. A number of works survey the region's long historical significance, including EMIL LENGYEL, *The Danube* (1939); JOSEPH WECHSBERG, *The Danube: 2000 Years of History, Myth, and Legend*

(1979); and SPIRIDON G. FOCAS, *The Lower Danube River in the Southeastern European Political and Economic Complex from Antiquity to the Conference of Belgrade of 1948,* trans. from Romanian (1987). (P.G.P.)

Materials in English on the Elbe, Oder, and Vistula rivers are scarce. Two brief works on the Elbe are K. SCHMIDT, "Hydrological Structure of the Federal Republic of Germany," pp. 31–39 in HANS-JÜRGEN KLINK and HERBERT LIEDTKE (ed.), *Physical Geography in the Federal Republic of Germany* (1984); and G. LUTTIG and K.-D. MEYER, "Geological History of the River Elbe, Mainly of Its Lower Course," pp. 1–19 in P. MACAR (ed.), *L'Évolution Quaternaire des bassins fluviaux de la mer Nord méridionale* (1974). The Elbe's regime is discussed in FRANKDIETER GRIMM, "Das Abflussverhalten in Europa, Typen und regionale Gliederung," *Wissenschaftliche Veröffentlichungen des Deutschen Instituts für Länderkunde* 25/26: 18–180 (1968). A.C. SEMMLER (ed.), *Der Elbstrom, von seinem Ursprunge bis zu seiner Mündung in die Nordsee* (1845, reprinted 1984), is a comprehensive work. (F.G.)

The only substantial works providing comprehensive coverage of the Oder and Vistula are in Polish and include JULIUSZ STACHÝ (ed.), *Atlas Hydrologiczny Polski,* 2 vol. (1987), containing maps and tables; and ZDZISLAW MIKULSKI, *Zarys hydrografii Polski* (1965), which, though dated, is still considered the fundamental professional study. ANDRZEJ GRODEK *et al.* (eds.), *Monografia Odry* (1948), is the standard source for the Oder. Useful information is also found in DON E. BIERMAN, *The Oder River: Transport and Economic Development* (1973), focusing on shipping and navigation. Surveys of the Vistula include ANDRZEJ PISKOZUB (ed.), *Wisła, monografia rzeki* (1982); and ALEKSANDER TUSZKO, *Wisła przyszłości* (1977). LESZEK STARKEL (ed.), *Evolution of the Vistula River Valley During the Last 15000 Years,* trans. from Polish, 2 vol. (1982–87), explores the geomorphology of the area. JAN STYCZYŃSKI, *Vistula, The Story of a River* (1973; originally published in Polish, 1973), is a descriptive pictorial work. JAN CZARNECKI, *The Goths in Ancient Poland: A Study on the Historical Geography of the Oder–Vistula Region During the First Two Centuries of Our Era* (1975), is a concise examination of events in relation to the geographic setting. (Je.P.)

(*Eastern European drainage systems*): The Dnepr, Don, and Volga rivers are often treated together because of their physical and economic interaction. Survey information is found in such general sources as NATIONAL GEOGRAPHIC SOCIETY, *Great Rivers of the World* (1984); MICHAEL T. FLORINSKY (ed.), *McGraw-Hill Encyclopedia of Russia and the Soviet Union* (1961); S.V. KALESNIK and V.F. PAVLENKO (eds.), *Soviet Union: A Geographical Survey* (1976; originally published in Russian, 1972); and, in Russian, M.I. LVOVICH, *Реки СССР* (1971). The following works study the influence of civilization and human interference on riverine biology, ecology, and river flow: I.A. SHIKLOMANOV, *Антропогенные изменения водности рек* (1979); S.L. VENDROV, *Проблемы преобразования речных систем СССР,* 2nd rev. ed. (1979); A.B. AVAKYAN and V.A. SHARAPOV, *Водохранилища электростанций СССР* (1962), focusing on water reservoirs and hydroelectric power plants; F.D. MORDUKHAI-BOLTOVSKOI (ed.), *The River Volga and Its Life* (1979; originally published in Russian, 1978), on the flora and fauna of the Volga and their changes; PHILIP P. MICKLIN, "Environmental Costs of the Volga–Kama Cascade of Power Stations," *Water Resources Bulletin* 10(3):565–572 (1974), and a longer article by the same author, "International Environmental Implications of Soviet Development of the Volga River," *Human Ecology* 5(2):113–135 (June 1977); and S.L. VENDROV and A.B. AVAKYAN, "The Volga River," pp. 23–38 in GILBERT F. WHITE (ed.), *Environmental Effects of Complex River Development* (1977).

There are a number of writings describing travels along the Russian rivers and providing political and social insights: MARVIN KALB, *The Volga: A Political Journey Through Russia* (1967), originated as a television documentary; HOWARD SOCHUREK, "The Volga, Russia's Mighty River Road," *National Geographic* 143(5):579–613 (May 1973), reports a trip by an experienced journalist; and DANIEL R. SNYDER, "Notes of a Visit to the Middle Volga," *Soviet Geography* 21(3):180–183 (1980), describes a cruise on the Volga and Don and visits to the major cities of the area. The many relevant historical works include the following: RICHARD G. KLEIN, *Man and Culture in the Late Pleistocene* (1969), which deals with the Stone Age civilization of the Don River valley; BORIS A. RAEV, *Roman Imports in the Lower Don Basin,* trans. from Russian (1986), based on the result of the archaeological excavation in the Don River region; ROBERT PAUL JORDAN, "Viking Trail East," *National Geographic* 167(3):278–317 (March 1985), which explores the role of the Volga and Dnepr in the founding of the early Russian state; ELVAJEAN HALL, *The Volga: Lifeline of Russia* (1965), a concise historical work; WILLIAM T. ELLIS, "Voyaging on the Volga amid War and Revolution: War-Time Sketches on Russia's Great Waterway," *National Geographic* 33(3):245–265 (March 1918), which focuses on the events of the first two decades of the 20th century; MAYNARD OWEN WILLIAMS, "Mother Volga Defends Her Own," *National Geographic* 82(6):793–811 (December 1942), which explores life along the Volga in the period before World War II; ANNE D. RASSWEILER, *The Generation of Power: The History of Dneprostroi* (1988), which surveys the construction of the power plant on the Dnepr; and BORIS SHIROKOV (comp.), *The Undying Tradition: Folk Handicrafts in the Mid-Volga Region,* trans. from Russian (1988), which explores the cultural tradition influenced by the great Russian rivers. (P.P.M.)

European History and Culture

The history of Europe is the story of the growth, development, and spread of a distinctive culture from its origins within a small geographical area at the western end of the Eurasian landmass. Economic, political, and social events influenced the development of that culture, and periods of harmony alternated with periods of discord among the various European peoples. From very early times, however, Europeans showed the determination to spread their civilization beyond its geographic home.

The Hellenistic and Roman empires both influenced and were influenced by the peoples they conquered and ruled. Northern European tribes (the so-called barbarians) migrated across the continent in the early centuries AD, and the meeting of northern and Mediterranean peoples resulted in a new, distinctive culture. "Christendom" was born through the spread of a new religion, and the expansion of the Christian faith was accompanied by the spread of the new culture. Later explorations were undertaken for economic and political reasons as well as continuing missionary efforts, but these too resulted in the enrichment and increasing influence of European civilization (see also EUROPEAN OVERSEAS EXPLORATION AND EMPIRES, THE HISTORY OF).

Economic resources, favourable geographic and climatic conditions, and the size of the population were among the factors that allowed this civilization to flourish. (For a discussion of the physical and human geography of the continent, see EUROPE.) By the beginning of the modern period, Europeans had spread their society and institutions throughout the world, and North and South America, Australia, and New Zealand were largely European in cultural tradition. The successful spread of the culture led, however, to the development of societies that in the 20th century began to compete with Europe for economic power and political influence. (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 921, 923, 961, 962, 963, 971, and 972.

This article is divided into the following sections:

Prehistory

SPREAD OF NEOLITHIC FARMING COMMUNITIES

Asian and indigenous influences

Agriculture and stock raising, the economic basis of peasant and ultimately of urban life, first appeared in the Aegean about the end of the 7th millennium BC but did not reach the more remote parts of temperate Europe for another 2,000 or in places even 3,000 years. Although settled life based on farming had been developed earlier in parts of southwestern Asia, Europe played a far from passive role. Horses, for instance, were almost certainly domesticated first in southern Russia and dogs in northern Europe; and it may even be that cattle were first tamed in southeastern Europe. Again, while it is true that many elements of material culture, notably pot making and the equipment used in harvesting agricultural crops, were invented earlier in southwest Asia, there was certainly no question of the spread of a single, uniform culture over the European continent. The development of Neolithic farming communities, which lay at the very basis of European civilization, was as complex as it was prolonged, and the diverse pattern to be found in the archaeological record reflects considerable variety in the culture of the indigenous Mesolithic population.

First phase. Communities of mixed farmers appeared in mainland Greece and on Crete about 8,000 years ago. Their most visible memorials are the mounds, or tells, formed, as in southwest Asia, by the disintegration of settlements of mudbuilt houses rebuilt generation by generation on permanently occupied village sites. The artifacts from such settlements—the handmade pots, clay figurines of women, stamps or seals, reaping knives, querns, and polished stone axes—resemble in their general form and character those from Syria, Iraq, and Iran. Yet the pots from Greece have their own idiosyncrasies of style. The vessels with simple, geometric, painted designs that accompany plain wares from this period in Thessaly or western Macedonia differ from comparable Asiatic wares and contrast even more strongly with the deeply incised wares from early levels at Knossus in Crete, whose affinities lie rather in Anatolia.

Second and third phases. By the middle of the 6th millennium permanent tell settlements had begun to grow up in the southern Balkans. Pottery and ancillary objects in the Starčevo-Körös style occur with minor local variations over a territory extending from the Adriatic coast of Yugoslavia, across to Bulgaria and Romania, and as far as the Prut River in present-day U.S.S.R. The main characteristics of this extensive culture are common to those earlier established in Greece, and it is significant that some of its distinctive features, such as the prevalence of surface roughening or rustication and the occasional use of naturalistic plastic ornamentation, appear on pottery from Greek Macedonia. The middle of the 5th millennium witnessed the expansion of farming economy over a territory centred on middle Europe, but extending to the Rhineland, the Low Countries and northern France in the west, to Hannover and Silesia in the north, and in the east outflanking the Carpathians as far as the western Ukraine. This culture, called Danubian, used pottery ornamented by pairs of parallel lines arranged in spiral or meander patterns. The Danubian peasants lived in communities very similar to large modern villages in the same territories. Their rectangular farmhouses, commonly up to 30 or even 40 metres long, were built on massive timber frames and apparently included storage space and possibly stalls. Their settlements were not marked by tells as those of the Starčevo-Körös and Aegean culture areas. This was due in part to the fact that it was no longer practicable in more temperate latitudes to build in sun-dried brick or *pisé* but in part also to the fact that the first peasants to occupy the loess of middle Europe appear to have practiced a form of shifting agriculture that involved frequent changes of settlement. In addition to practicing the same form of economy, living in the same sorts of houses, and making the same kind of pottery, the Danubians everywhere favoured polished stone adz blades for the working of timber, buried their dead in crouched positions in inhumation cemeteries and, at least in the later stages of their culture, were careful to provide them with ornaments that were made from the *Spondylus gaederopus* mussel.

Danubian culture

Fourth phase. The fourth phase in the expansion of farming economy in Europe can be studied in three main areas. In the eastern area a new peasant culture, the Boian, grew up in Romania beyond the limits of the Starčevo-Körös area; it used distinctive forms of pottery that were decorated by surface rippling, painting, and by excised or incised linear designs. To the north of the Boian culture, and owing something to its inspiration as well as to the inspiration of the Danubian, another one, called after Cucuteni and Tripolye, developed between the Seret and the Bug rivers with an extension to the Dnepr. The pottery of this culture was fired in well-controlled ovens to a red or orange colour after having been decorated in curvilinear designs that were painted onto or grooved into the surface. Its makers occupied rectangular houses that they grouped into substantial villages. In some of these villages the houses were arranged radially. Kolomiishchina, near Kiev, comprised a circular setting of about 39 houses with others in the middle; another at Vladimirovka, also near Kiev, had at least 162 houses in five concentric rings.

On the North European Plain during the 4th millennium, it appears that the hunter-fisher population was influenced by the Danubian peasants to the immediate south. The immediate outcome was the Ertebølle-Ellerbek culture, based on a mixed hunter-farmer economy and marked in the archaeological record by pointed based pottery and specific forms of flint and antler artifacts. The end of the

The Funnel-Neck Beaker industry

4th millennium saw the emergence of the Funnel-Neck Beaker industry, a farming economy extending from what is now The Netherlands across Poland and as far north as southern Norway and middle Sweden.

The Megalithic cult. At approximately the same time, a distinct cultural province developed in western Europe over a territory which included much of Switzerland, France, the British Isles, and Iberia. Excavation of the waterlogged deposits of the Swiss lakeside settlements belonging to the Cortaillod culture have yielded particularly full information about subsistence and crafts, such as weaving and embroidery, generally absent from the archaeological record. More significant is the region's contribution to the spiritual life of the time, as regards both funerary ritual and public ceremonial activities. The most numerous visible monuments dating from the 3rd millennium and the centuries either side of it are chamber tombs, sometimes cut out of the living rock, more usually built from megalithic blocks packed around with smaller stones, and held in place by overlying mounds. The construction of such tombs, designed as burial places for successive generations of families or clans, involved a physical and social effort comparable to that implied by the building of churches in early historic times and implies well-disciplined religious emotion. Some insight into the nature of the Megalithic cult is provided by the symbolic designs pecked, carved, or painted on stones incorporated into megalithic structures. The spread of this cult can hardly be accounted for in simple terms of diffusion, still less of wholesale ethnic movements. The variety of tomb forms and the varying form and incidence of symbolic art point to intense local and regional development and the transmission of ideas over a rather long time. It is significant that by and large the grave goods relate to the indigenous cultures.

In addition to funerary structures, these people erected single, standing stones, or menhirs, alignments and other elongated structures, and circular monuments, sometimes on a massive scale. The most impressive alignments are

Menhirs

those at Carnac, Brittany, comprising nearly 3,000 menhirs in parallel rows over a distance of between three and four miles. Smaller alignments occur in southwest England. The British Isles have many circular settings of menhirs or heavy timbers dating from this period; the best known example is Stonehenge.

The closing centuries of the 3rd and the opening ones of the 2nd millennium were marked by the widespread adoption over much of north, central, and eastern Europe of perforated stone battle-axes. This was during the time when farming economy spread in European Russia far beyond the forest steppe (to which the Tripolye peasants had been confined) over the southern Taiga up to the Oka-Volga basin, the homeland of the Fatyanovo culture. In northern Europe the Funnel-Neck Beaker culture gave place to a variety of local battle-ax-using groups all associated with single-grave burial, but in other respects showing many differences. Here, again, it seems that mixed farming continued to provide the basis of subsistence, and there is evidence that a light plow was already in active use as well as disk-wheeled wagons which likewise were drawn by yoked oxen. The construction of collective tombs was abandoned about the same time in western Europe. The

The Bell-Beaker culture

rapid and widespread occurrence over large parts of central and western Europe of pots of bell-shaped beaker form, decorated in horizontal zones by finely toothed stamps, suggests the movement of small mobile groups; and there is evidence that these actively sought out copper and gold.

The circumpolar culture. North of the farming zone, extensive parts of circumpolar Europe continued to support communities that lived by hunting, fishing, and gathering. Although there was some overlapping and contact, the culture of the circumpolar people was quite distinct from that of the peasant peoples of the rest of Europe. Even the art of potting, which they took over from their neighbours, was used to make distinctive ovoid pots reminiscent of basket forms; and the deep pits, which together with stamped comb imprints served to decorate these vessels, themselves stemmed, like much else in the culture, directly from Mesolithic antecedents. The widespread use of polishing was another archaic element originally developed in relation to bone but now applied to a variety of knives, projectile heads, and hollow-ground adzes of slate and stone. As hunters the people were accustomed to movement. During the extensive period of snow cover they used heavy sledges, presumably drawn by dogs, for shifting camp and skis for hunting and personal movement. Skin-covered boats represented on rock engravings were probably used for hunting seals and porpoises and for fishing. Dependence on hunting helps to account for the keen observation shown in the representation of animals that played a leading role in their graphic art. Engravings, sometimes life-size, of game animals and carvings of animal heads on a variety of artifacts, from perforated stone axes to the handles of wooden ladles, reveal a mentality totally opposed to that of the peasants to the south, whose art was mainly abstract and repetitive. Burial rites were also distinctive, taking the form of interment in cemeteries, the body fully clothed and in an extended position.

THE BRONZE AGE (C. 2500–C. 650 BC)

The beginnings of European metallurgy

While the peasants who pioneered farming economy in remote territories of the temperate zone were restricted to stone, metallurgy was already being practiced in parts of southeastern Europe and Iberia. The smelting of metallic copper from its ores was in itself a remarkable feat in the history of technology; it marked the first of what was to prove a long series of operations by which men were able to conjure up new materials, and it is worth emphasizing that to obtain ore, smelt it, and convert it into artifacts implied a more advanced degree of specialization of labour than was needed for a technology based on working flint or stone. It is no accident that in the Old World metallurgy should have been one of the features persistently found in the context of the earliest literate civilizations. Indeed, it has commonly been assumed that the appearance of metallurgy in prehistoric barbarian societies can only be explained in terms of their contacts with early centres of civilization. In the case of southern Iberia, eastern Mediterranean and specifically Cycladic (south Aegean) influence has been detected at copper-working settlements dating from the mid-3rd millennium; the defenses of Los Millares, for instance, display bastions identical in pattern with those of Chalandriani on Syros (Greece). Here it looks as though there were "factories" established for the prime purpose of securing supplies of copper for Aegean markets. The initial impetus for opening up the copper resources of Transylvania may also have come from the demands of Aegean and possibly even Asiatic markets. It is no less true that the exploitation of copper ores in southeastern and central Europe underwent its own development. Whereas in Iberia the initial impetus was soon lost, the opening up of ores in Transylvania led to the development of cuprous industries, first in Bohemia, central Germany, and the east Alps, and ultimately as far away as Ireland, where copper deposits were first sought out by the Bell-Beaker prospectors at about the end of the 2nd millennium. These copper resources catered wholly or in large measure to barbarian markets. Thus shafthole axes and ax-adzes of Transylvanian copper are found predominantly in Transylvania and the Danubian basin both above and below the Iron Gate, though a few reached the Elbe basin; spiral armbands and embossed disks reached as far north as Denmark, where they occur at the end of the local Early Neolithic. Again, flat axes of trapezoidal plan of the form made from east Alpine and central German copper found their way as far north as Scania (Scandinavia).

The earliest bronze industries

Central Europe was also the original focus of the local use of bronze alloys that not only resulted in tougher tools, but, by making it easier to use valve molds, made it possible to produce more complex forms. The new industries—those of Únětice on the Upper Elbe, of Tomaszów in Poland, of Kistapostag and Perjámos in Hungary, of Straubing on the Upper Danube, and of Adlerberg on the Middle Rhine—produced a variety of bronze items, including daggers, neck ornaments, and pins for securing garments. The occurrence of similar ornaments at Byblos in Lebanon and Ugarit in Syria dating from *c.* 2000–1800 BC suggests continued contact between middle Europe

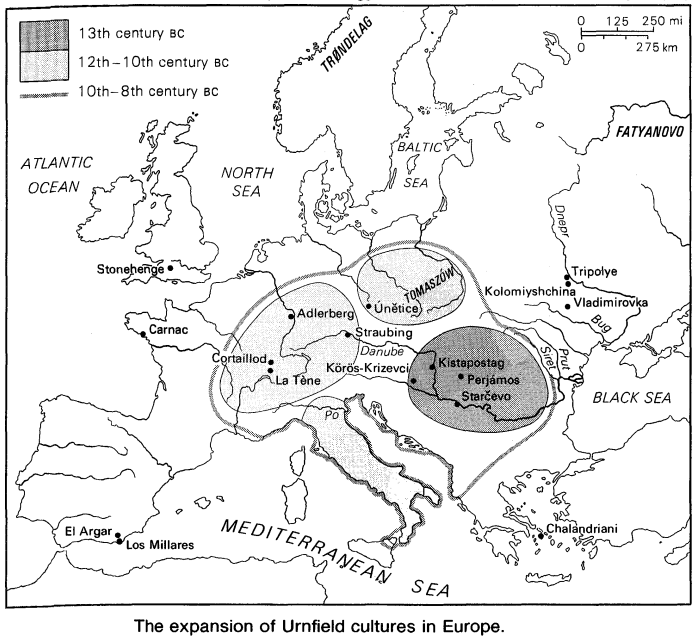
and areas of higher culture to the southeast, but this does not mean that one must attribute the advanced bronze industries of middle Europe around the middle of the 2nd millennium to stimulus from this direction. The Únětice and allied metal industries rested on a middle European tradition already many hundreds of years old. Again, regional bronzeworking traditions were developing at this time in many parts of Europe as far apart as Spain, where metallurgy revived in the El Argar culture, and Britain and Ireland, which gave birth to a distinctive Hiberno-British style.

The use of tin as an alloy laid a great emphasis on organized systems of exchange, since this metal was one with a more restricted distribution than copper. These involved the movement not only of ingots but of finished artifacts, as may be seen in the importation of Hiberno-British and middle European forms of ax to Denmark during the final phase of its Stone Age, a period significantly marked above all by flint daggers based ultimately on metal ones too expensive to import in adequate numbers. It is interesting to note how the Aegean and by this time the specifically Mycenaean world became enmeshed in this wide-ranging system of exchange, due in all probability to their need for central European or Cornish tin. The occurrence of segmented faience beads, of a kind invented in Egypt but probably made also in the Aegean, in the territory of the Únětice and allied cultures, and also notably in Wessex, a territory that probably controlled supplies of Cornish tin, is suggestive of Mediterranean influence in the north. But the importance of the Mycenaeans in the European Bronze Age should not be overemphasized. The indirect effects of the play of the Mycenaean market may have been important—for instance, the Mycenaean demand for amber probably played its part in the rise of the Nordic Bronze Age by making it possible for the Danes to import central European bronzes and ingots—but metalsmiths in several parts of barbarian Europe were producing bronzes that compare favourably with anything produced in the Aegean at the time; and faience beads of local shapes were produced even in northern Britain.

The loss of Mycenaean markets was no longer enough to affect the prosperity of central Europe. Copper was mined on an increasing scale and, being cheaper, was used for tools, armour, and containers, as well as for weapons and ornaments. The later stages of the Bronze Age also were marked by a greater emphasis on cremation and the burial of ashes in urns. Urnfield burial, together with many innovations in industry and armament, was adopted over extensive tracts of Europe from Iberia to eastern Europe and from Italy to southern Scandinavia. How much this was due to the normal process of diffusion by which innovations spread over extensive territories in the course of time, and how much to ethnic movement, is still obscure. Yet it is hard not to believe that raids by armed marauders did not play a part more especially in relation to the rich territories of the Mycenaean world. Already in the 15th century BC heavy slashing swords, with flanged grips to secure the handle, were being made in east central Europe, and such swords were among the innovations that spread widely with the adoption of urnfield burial. The sudden appearance of these in mainland Greece, Crete, and Cyprus has been identified by some scholars with intrusion f the Dorians mentioned by Thucydides as entering the Peloponnesus at a time when archaeology speaks of the destruction of Mycenaean civilization. Egyptian records tell of raiders from the north as early as the 13th century BC and, in the ensuing centuries, the east Mediterranean region as a whole underwent a time of troubles, a period that witnessed the decline not only of the Mycenaean but also of the Hittite power. Other peoples were on the move at this time and may well have been more important, but it looks as though pressure may have come in part from the north. (J.G.D.C./Ed.)

Urnfield cultures

Adapted from S. Piggott, *Ancient Europe* (1965); Aldine Publishing Company



The expansion of Urnfield cultures in Europe.

The "barbarian" peoples

The interrelated cultures of ancient Europe that were fully established by about 1000 BC incorporated indigenous traditions and stocks, from region to region, but from this time the archaeological evidence in cultural zonation leads on ever more clearly to the appearance in history of the European barbarian peoples.

CELTS

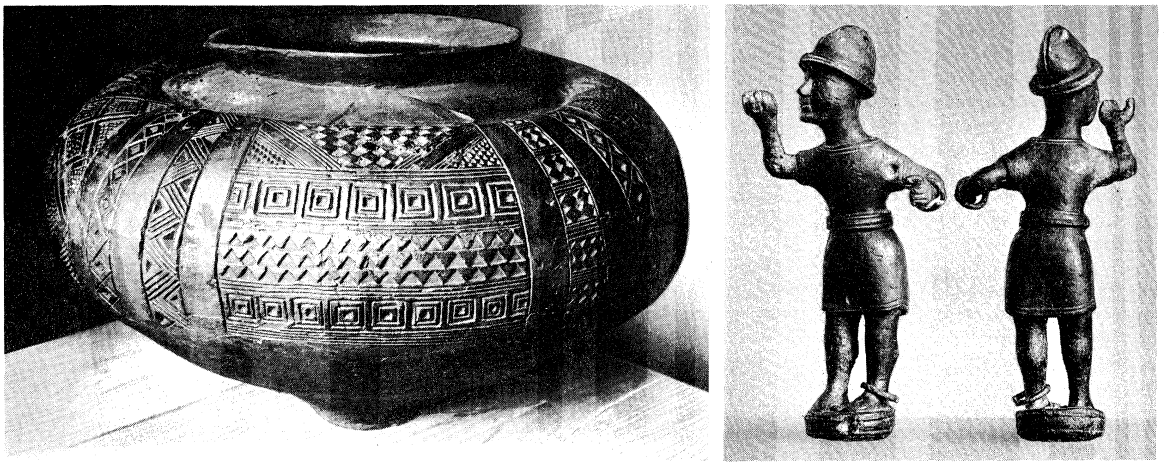
The Celts were the first prehistoric people to rise from anonymity in the European territories north of the Alps. Until the middle of the first millennium BC nothing was known of them by the civilized Mediterranean world. For the Greeks they were *Keltoi;* the Romans called them Gauls. At the time when these Celts had become the predominant people in the barbarian world they were settled throughout a great part of Europe, extending from Ireland and Britain to the Balkans, and even as far as Anatolia. They were a numerous people of considerable political and military significance. Their creative activity completed the prehistoric development of Europe; by the 4th century BC Greek writers ranked the Celts, together with the Scythians and the Persians, among the most numerous "barbarian" peoples of the then known world. Although they never formed a unified ethnic group or a great empire and were split into many tribes with different dialects, they became an important factor in the development of a specifically European civilization.

Early Celtic settlements. Archaeological investigation strongly suggests that the area in which the historical Celtic tribes developed comprised part of present-day France, southern Germany, and adjacent territory reaching as far as southern and central Bohemia. It is possible to trace their origins as far back as the Bronze Age Tumulus culture, which reached its high point about 1200 BC. Various cultural components were unified in the succeeding Urnfield culture (12th–8th centuries BC). This was followed by the Hallstatt period (7th–6th centuries; named after the burial ground near Hallstatt in the Salzkammergut in upper Austria). During that time other peoples known to history made their appearance in Europe: the Illyrians in the southeast and the Germans in the north.

From the Urnfield period on, archaeological evidence of the Celtic confederation northwest of the Alps is increasingly noteworthy, producing, in the later Urnfield stage, a fresh phase of Tumulus burials, some of these being located in areas where Celtic power later became predominant.

Burial finds

Burial grounds of the Hallstatt period comprise chamber graves containing four-wheeled wagons, splendidly decorative yokes and harness fittings, and a large amount of pottery and other artifacts, which indicate a complex burial ceremony. Such elaborate burials, found in Bohemia, throughout southern Germany, and in northeastern France, would appear to be those of chieftains or



Artifacts of Celtic cultures.
(Top left) Pottery vessel, Hallstatt culture, 7th–6th century BC. In the Württembergisches Landesmuseum, Stuttgart, West Germany. Height 14 cm. (Top right) Bronze warrior, Hallstatt culture, 6th–5th century BC. In the Naturhistorisches Museum, Vienna. Height 7.9 cm. (Bottom left) Bronze flagon decorated with coral, La Tène culture, late 5th century BC. In the British Museum. Height 38.8 cm. (Bottom centre) Back view of a mirror, La Tène culture, early 1st century BC. In the British Museum. Height 35 cm. (Bottom right) Bronze enamelled shield, La Tène culture, early 1st century BC. In the British Museum. Height 77.5 cm.

By courtesy of (top left) the Wurttembergisches Landesmuseum, Stuttgart, (top right) the Naturhistorisches Museum, Vienna, (bottom left, bottom centre, bottom right) the trustees of the British Museum

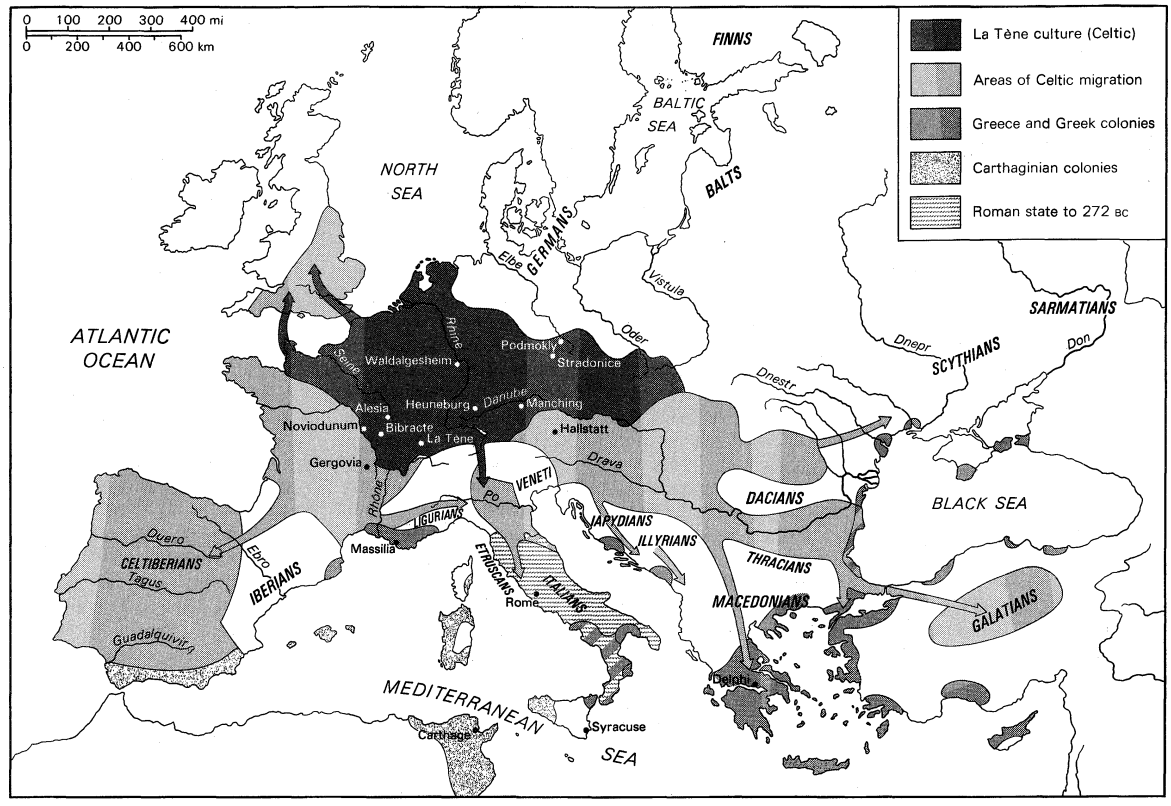
members of a leading oligarchy. To this period there may also be ascribed the first known fortified castles of western central Europe. The fortified hill residence at Heuneburg on the Danube, dating from the late Hallstatt period, was rebuilt at least five times. One phase shows a mature building technique including the use of brick masonry in towerlike bastions, a Mediterranean characteristic quite alien to the region. Adjacent to Heuneburg is a group of chieftains' barrows, one of which, Hohmichele, has wood-lined chambers and wall draperies and contains a four-wheeled wagon, gold decorations, and jewelry. It is clear that imports from the south, such as amphorae of wine or Greek Black Figure pottery, were common. A fortified hill site on Mt. Lassois near Châtillon-sur-Seine, France, in which Greek pottery was found, was built at the junction of long-distance trade routes. A bronze vessel, also of Greek manufacture, was found in the neighbouring grave of a Celtic princess.

Evidence of wealth

The wealth of the Celtic princes seems to have been very great, some of the barrows (mound graves) being particularly richly furnished for their size. That at Apremont (Haute-Saône), with an average diameter of 70 metres, contained a wagon, gold objects, a crown, brooches, and a vessel. The grave at Pflugfelden near Ludwigsburg (Württemberg) was the burial of a man in a splendid robe and a golden crown; the hilt of his dagger was inlaid with amber. By the 6th century the wealth of the ruling Celtic groups was well established and they were able to import Etruscan wares, the merchants of Massilia (Marseille) acting as distributing agents.

The unusual concentration of early Hallstatt chieftains' burials on the upper Danube and on the upper Rhine as far north as the Neckar River lasts only to the beginning of the 5th century. At that time the Celts were preparing armed raids into the remaining parts of Europe. Thereafter, rich burials are found farther northwest, on the middle Rhine or on the Saar and Moselle rivers. The number and the richness of the burials steadily increased, including two graves at Schwarzenbach that contain two beaked flagons, a vessel in a gold openwork casing, and various gold ornaments showing mask motifs.

The La Tène culture. Artifacts from the La Tène culture, which developed in the 5th century BC, are found in the barrows of the middle Rhine and of Champagne in France. From the second half of the century at the latest native Celtic craftsmen were adapting southern models and developing their own style with increasing confidence. The geometric patterns found in ornament of the Hallstatt period give place to zoomorphic and floral designs, supplemented by human-head motifs or masks.



The Celtic migrations.

From Grosser Historischer Weltatlas, vol. 1, *Vorgeschichte und Altertum* (1972); Bayerischer Schulbuch-Verlag, Munich

A remarkable genre of the style are the bronze-mask brooches, adorned with human and animal masks varying from lifelike to fantastic portrayal. Such brooches have been found at Panenský Týnec near Louny (animal mask) and near Manětín-Hrádek in the Plzeň region (imaginative human likeness) in Czechoslovakia and at Parsberg and Oberwittighausen in Germany.

These early La Tène pieces, typical of the second half of the 5th century BC, are still much influenced by southern models. The mature La Tène style, also called the Waldalgesheim style, dates from the second half of the 4th century, and its products are found in chieftains' graves throughout Europe. In the 3rd century the use of jewelry extended to all social classes, and craftsmen became more numerous, no longer working exclusively in the entourage of the rulers. While elaborately linked bracelets and brooches became common ornaments, the skill of leading craftsmen was adapted to ornamenting swords, scabbards, and helmets for the aristocracy.

The development of Celtic art in Britain was quite individual in character. Local workshops and "schools" were established, in which the native craftsmen combined technical skill with a freshness of conception. When the Celtic art of the Continent had passed its zenith, that of Britain continued to flourish, producing a variety of beautiful artifacts, such as gold and silver jewelry, swords and scabbards, shields inlaid with enamel, and bronze mirrors.

Celtic expansion. During the Hallstatt period, the Celts expanded through France to the Iberian Peninsula, to the British Isles, and also to some extent eastward into central Europe. Military expansion was probably due to overpopulation and social tensions. Celtic bands also entered Italy, groups such as the Boii, the Insubres, the Lingones, and the Senones first attacking Etruria, while later groups reached the Adriatic coast and about the year 387 BC (the conventional date is 390) raided and plundered Rome, also penetrating into southern Italy. The Romans later recovered and drove the Celts back to the Alpine foothills, whence some of them withdrew, probably into central Europe.

In the second stream of expansion their raids reached central Europe, the Carpathians, and the Balkans. This expansion was later described by the Roman historian Livy, who recounts how two branches of the Bituriges settled near the Hercynian Forest, a range of mountains of southern and central Germany (under their ruler Sigovesus) and in Italy (under Bellovesus). Literary sources suggest that the Celts reached the Carpathians in the 4th century, later moving into present-day Bulgaria, Romania, Thrace, and Macedonia; they raided the shrine at Delphi, Greece, in 279. One group reached Anatolia, where they settled and gave their name to Galatia. In the 3rd century Celts were serving as mercenaries in Greece, Anatolia, and Egypt.

Incursion of the Cimbri

An important dividing line in the history of the Celts was the incursion of the Germanic Cimbri. Their assault was repulsed by the Boii about 113 BC, somewhere in the neighbourhood of Bohemia. The Cimbri then attacked the Scordisci near Belgrade and later, in association with the Teutoni, attacked several Celtic tribes in the west, penetrating as far as Aquitania in Gaul. Although the Cimbri were defeated by the Romans in the Po Valley at the end of the 2nd century, their incursion into Celtic territory foreshadowed later invasions.

Gaul and Britain. In the 1st century BC Gaius Julius Caesar described Gaul as having three regions, inhabited respectively by the Galli (Gauls, present Central France), the Aquitani (southwestern France), and the Belgae in the north. The Gauls comprised many tribes: the Helvetii in present-day Switzerland, the Sequani and Lingones farther west, the Arverni in Auvergne, the powerful Aedui between the Saône and Loire rivers, and the Bituriges along the Loire Valley. The Celts reached the shores of the Mediterranean later than other parts of Gaul; the Carthaginian commander Hannibal made contact with them there late in the 3rd century BC. In the 1st century, Caesar subdued the Helvetii and other tribes, pressed the Germans back beyond the Rhine, and made two expeditions to Britain. Finally, he suppressed a revolt of Gallic tribes under Vercingetorix. From that time, Gaul was subject to Rome and increasingly open to Romanization. Britain was subjugated in the 1st century AD.

In the 2nd century BC the Romans reached southern Gaul and founded their province Gallia Narbonensis (Provence).

The names of many tribes in the British Isles are known: the Dumnonii in Cornwall, the Dobuni on the upper Thames, and the Ordovices in Wales. About the middle of the 3rd century BC a new wave of immigrants from Gaul introduced the La Tène culture into Britain. Some Belgic tribes (the Cantae, Catuvellauni, Atrebates, Durotriges) settled in Britain in the 1st century BC.

The final phases. From the middle of the 1st century BC the Celtic world was caught in the press of two dangerous forces: the Roman Empire was extending its frontiers to the Rhine and the Danube, while the Germanic tribes were thrusting southward. Even before this time a strong settlement of the Celtic Boii existed in Pannonia, where they had apparently taken refuge after being driven out of Bohemia. After Burebistas, king of Dacia, destroyed the Boii (*c.* 50 BC), the Celts withdrew across modern Switzerland into Gaul. By the end of the century the Celts on the Continent had lost their commanding position and the Rhine and the Danube and were confined to the frontier between the Roman and Germanic worlds.

The Celtic tradition survived most strongly in Britain and Ireland, where its art was still flourishing at the beginning of the Christian Era. Several of its elements were revived in the neo-Celtic art forms of the 7th to 9th centuries AD.

Celtic society. In the eyes of Greeks and Romans, the Celts were remarkable for their height, muscularity, and fair colouring. These were characteristics of the warrior class rather than of the whole population, and skeletal remains point to considerable variations in stature and head form. The 1st-century-BC Greek geographer and historian Strabo describes them as a people who love war and adventure, pleasure and feasts. Information on Celtic institutions is available from various classical authors and from ancient Irish literature. A striking feature, even as early as the La Tène period, is the large number of tribes competing for dominance. Kingship was common among some tribes in the La Tène period. Some kings, however, were elected, and the power of the aristocracy tended to increase, their rule predominating in Gaul in Caesar's time. In Ireland, by customary law, the social system was threefold; king, warrior aristocracy, and freemen farmers. In the later period in the western Celtic world, the strength of the aristocracy was increased and the freedom of common people was diminished by the widespread introduction of clientism, a system by which lesser men placed themselves under obligation to a powerful lord from whom they received protection. As in other Indo-European systems, the family was patriarchal and kingship was recognized by agnatic descent (from a common male ancestor). Landownership was vested in the family, which was also responsible for many social obligations. The household was of the "archaic-joint" type, consisting of a man with his wives, children, and grandchildren. The status of women seems to have varied considerably according to rank and the prosperity of the community.

Druids

The Druids, who were occupied with religious and legal duties, were recruited from families of the warrior class but ranked higher. They offered sacrifices and were responsible for the education of the young nobles, and the sacred oak groves were under their protection. Druidism, still flourishing at the beginning of the Roman occupation of Celtic areas, strongly resisted Romanization and was thus particularly marked out for deliberate suppression. It survived longest in Ireland, where the Romans did not penetrate. Caesar describes the Celtic sacred sites, woods with sacred trees and groves (*loci consecrati*). There is no evidence that they had temples in his time, but these were built later, in the Gallo-Roman period. Cult sites have been found, however, in several places in central and western Europe. An example is at Libenice near Kolīn in Bohemia, excavated in 1959, which had a stone slab for sacrifices. In the centre of the cult area a woman was found buried, with typical La Tène ornaments. Caldrons connected with the cults spread as far as northern Europe, and splendid examples have been found in silver at Gundestrup and at Brå in East Jutland. It is difficult to generalize about Celtic religion and mythology, because each tribe seems to have had its own cults and local deities. There is no firm distinction between gods and heroes; the names of some deities (*e.g.*, Taranis, Teutates, Esus, Lug) antedate the Romanization of Gaul, but later they sometimes became associated with Roman gods. Divine or semidivine pairs and trinities also occurred.

Celtic economy

The basic economy of the Celts was mixed farming, and, except in times of unrest, single farmsteads were usual. All kinds of grain were cultivated, also flax, hemp, beetroot, and several other kinds of vegetables. The grain was stored in special pits or silos deep in the ground or in large jars. Owing to the wide variations in terrain and climate, cattle raising was more important than cereal cultivation in some regions. The prevalence of oak forests encouraged the breeding of pigs, and boar hunts were a popular sport. The wool of Celtic sheep was well known in Rome. Later, Gaul supplied the Roman armies on the Rhine with horses.

Trousers, perhaps an Eastern innovation, were worn by men of the Cisalpine Gaulish tribes from at least the 3rd century BC and are attested in the 1st century BC in Transalpine Gaul, but a belted tunic, or shirt, with a cloak seems to have been the most widespread form of male dress. Women wore a single long garment with a cloak. Coarse linen as well as wool was employed, and bright colours were popular. According to Poseidonius the food of the Celts consisted of bread and meat, either boiled or roasted. Beer, home brewed from barley, was the most common drink; the upper classes also used wine imported from the south. The Celts were hospitable, fond of feasting, drinking, and quarrelling, and incapable of prolonged concerted action. At feasts bards sang the praises of those present, accompanying the song with a lyrelike instrument. The Celts greatly prized music and many forms of oral literary composition.

The earliest Celts had fortified strongholds or chieftains' residences; hill forts (*oppida*), with a considerable concentration of population and production, were constructed from the second half of the 2nd century BC, partly as a result of Germanic pressures. A belt of flourishing hill forts existed in Caesar's time, but his victory led to their decline in Gaul, although many in central Europe were occupied until the end of the last century BC. (J.F./Ed.)

GERMANIC PEOPLES

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

By the time of Julius Caesar, Germans were established west of the Rhine and had reached the Danube in the south. Their first great clash with Romans came at the end of the 2nd century BC, when the Cimbri and Teutoni (Teutones) invaded southern Gaul and northern Italy and were annihilated by Marius in 102 and 101. Although individual travellers from the time of Pytheas onward had visited Teutonic countries in the north, it was not until the middle of the 1st century BC that the Romans learned to distinguish precisely between the Germans and the Celts, a distinction that is made with great clarity by Julius Caesar. It was Caesar who incorporated within the frontiers of the Roman Empire those Germans who had penetrated west of the Rhine, and it is he who provides the earliest extant description of Germanic culture. In 9 BC the Romans pushed their frontier eastward from the Rhine to the Elbe, but in AD 9 a revolt of their subject

Germans headed by Arminius ended in the destruction of the occupying army of P. Quinctilius Varus in the Teutoburg Forest and in the withdrawal of the Roman frontier to the Rhine. In this period of occupation and during the numerous wars fought between Rome and the Germans in the 1st century AD, enormous quantities of information about the Germans reached Rome. When Tacitus published in AD 98 the book now known as the *Germania,* he had reliable sources of information on which to draw.

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

New tribes

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

By the year 500 further striking changes had taken place. The Angles and Saxons were in England and the Franks controlled northeastern Gaul. The Burgundians were in the Rhône Valley with the Visigoths as their western neighbours. The Ostrogoths were established in Italy and the Vandals in Africa. In 507 the Franks expelled the Visigoths from most of their Gallic possessions, which had stretched from the Pyrenees to the Loire River, and the Visigoths thereafter lived in Spain until their extinction by the Muslims in 711. In 568 the Lombards entered Italy and lived there in an independent kingdom until they were overthrown by Charlemagne (774). The areas of eastern Germany vacated by the Goths and others were filled up by the Slavs, who extended westward as far as Bohemia and the basin of the Elbe. After the 8th century the Germans recovered eastern Germany, lower Austria, and much of Styria and Carinthia from the Slavs.

Material culture. According to Julius Caesar the Germans were not primarily agriculturists; they were pastoralists, and the bulk of their foodstuffs—milk, cheese, and meat—came from their flocks and herds. But agriculture was not unknown. Grain and a great variety of root crops and vegetables were known to them, though of fruits only the apple was cultivated by them in the Roman period. As for the techniques of agriculture, Caesar and Tacitus are silent on these, but even the Bronze Age rock carvings at Bohuslän in southern Sweden include a picture of an ox-drawn plow.

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

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

Trade

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

Growth of the *comitatus*

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

Maroboduus' kingdom

siderable number of other Germanic peoples between the Elbe and the Vistula, including the Semnones, the Lombards, and the Lugii. But the Cherusci, joined by some of the king's subjects, attacked him in AD 17, overthrew him, and drove him into Roman territory. All other chiefs who attempted in this period to establish monarchies were, so far as is known, defeated.

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

The absence of monarchies

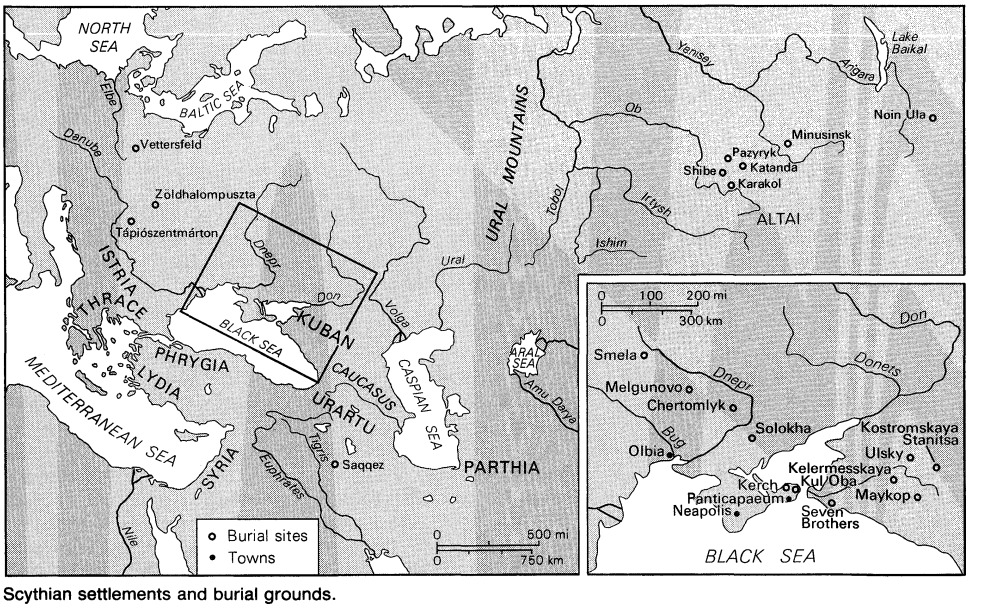
EASTERN PEOPLES

The grasslands that extended from the Black Sea northwestward outside the Carpathian Mountains as far as the Baltic Sea provided at the beginning of the 2nd millennium BC not only suitable habitation for pastoralists and mixed-farming communities but also means of long-distance intercommunication. Here Corded Ware people had first intruded on older communities. From the middle of the 2nd millennium BC various regional cultures have been distinguished, evidently representing continuity in settlement but showing varying degrees of influence from both the Bronze Age cultures of middle Europe and the steppeland. A phase of renewed activity over the steppe resulted in the appearance of the Cimmerians. Some burials and bronze horse trappings of the 9th century BC have been ascribed to them. Then, during the 8th and 7th centuries, followed the Scythians, whose warlike qualities and vivid animal art style had a profound influence even beyond the Carpathian Shield.

Scythians. *Migration and expansion.* During the 9th century BC the Scythian and kindred tribes were probably concentrated somewhat to the east of the Altai, but it was not until the Chinese ruler Hsüan Wang (827–781 BC) decided to send an armed force to curb the fierce Hsiung-nu, who had begun to make a practice of raiding China's western boundaries, that the Scythian nomads became restless. When the Hsiung-nu were forced back from the Chinese frontier and, in retreating, dislodged the Massagetai, who occupied the grazing grounds to the north of the Oxus (Amu Darya) River, and when the latter in their turn assaulted their immediate neighbours, the Scythians, a wide-scale nomadic migration was set in motion. There is reason for thinking that the struggle for grazing land was rendered more acute by a severe drought and that this factor may very well have convinced the Scythians to move westward rather than remaining to fight for their traditional rights. The Scythians were accomplished horsemen, among the earliest people to master the art of riding. Their mobility gave them a considerable advantage over their neighbours, so that when they eventually advanced westward across the Oxus they moved so rapidly that both Herodotus and contemporary Persian sources refer to the remarkable suddenness of their appearance on Iran's northeastern border. Their advance brought the Scythians into fierce conflict with the Cimmerians who had for centuries enjoyed possession of the Caucasus and the plain lying to the north of the Black Sea. The Cimmerians, however, still fought on foot and the Scythian cavalry quickly gained the upper hand. Some Cimmerians retreated through the Daryal (Dariel) Pass and were pursued across the Volga, where the Scythians were able to destroy and supplant them.

Destruction of the Cimmerians

Meanwhile another Scythian force chased the rest of the Cimmerian army across Urartu (Armenia), while a third entered the Derbent defile and reached Lake Urmia sometime during the reign of King Sargon of Assyria (722–705 BC), linking up there with the second contingent to continue the fight against the Cimmerians. Thus strengthened, the Scythian troops were able to force the Cimmerians into a steady retreat that lasted for about 30 years, ending only when both combatants had reached the borders of Assyria. Then the Scythians formed an alliance with King Esarhaddon of Assyria (reigned 680–669 BC), but they



Scythian settlements and burial grounds.

soon abandoned this and concentrated on destroying the Cimmerians, giving the latter no respite until they had forced them back across Phrygia into Lydia, where they were finally wiped out.

This astonishing series of victories brought great fame to the Scythians. Their chieftain Partatua (Bartatua) and his son Madyes were quick to take advantage of it by setting themselves up as rulers of west Persian lands stretching to the Halys (Kızıl Irmak) River, establishing their capital at Saqqez. They invaded Syria and Judaea *c.* 625 BC. Later they reached the borders of Egypt, but Psamtik I (663–610 BC) wisely decided to check any further advance by purchasing peace terms from them.

Meanwhile the Medes had become masters of Persia. They considered the Scythians' increasing might a real threat to their own security, and they decided to concentrate their efforts in launching a decisive attack against the tribesmen. Their better disciplined troops eventually contrived to push the Scythians northward, whence they had first appeared, some retreating through Urartu. Although the nomads thereby lost the control they had wielded during the previous 28 years over most of Anatolia, their lands still stretched from the Persian border through the Kuban into most of southern Russia. When forced by the Medes to retreat, some of them likewise settled between the Caspian and Aral seas, where they intermingled with their Dahae kinsmen to produce the people who, some three centuries later, were to become known as the Parthians, while others penetrated into India and established kingdoms there.

The Scythians in southern Russia

The Scythians who settled in the Kuban and southern Russia quickly attained a position of importance. Many of them became extremely prosperous. Graves of the 7th–6th centuries BC situated in the Kuban abound in objects of gold and other precious materials, and although a form of patriarchal rule remained in force, it is clear that a class of wealthy chieftains or nobles was beginning to come into being there. In the 6th–5th century BC the richest burials were in southern Russia and the Crimea. These are associated with a relatively small number of Scythians who, as the Royal Scyths, established themselves as rulers of the area. Isolated groups of their tribesmen penetrated as far as what became Hungary and East Prussia. The kingdom of the Royal Scyths developed into a community that was to enjoy considerable economic power until about the 1st century BC. Its political importance was established when it was able to resist Darius on his invasion of Scythian territory about 513 BC; the Scythians resorted to a scorched-earth policy, which enabled them to avoid a large-scale battle and obliged Darius to beat a hasty retreat in order to preserve his army. His consequent withdrawal from the plain lying to the northwest of the Black Sea left the Scythians in control of it; some of the Greek cities of the Pontus even had to pay the Scythians an annual tribute in order to preserve their security. Scythian power remained paramount there until the 4th century BC, when the Sarmatians appeared on the Don. Thenceforth the new invaders began steadily to increase their pressure on the Scythians until, in the 2nd century BC, they managed to confine them within the Crimean area, gradually supplanting them as rulers of the plain until, in the 2nd century AD, they finally succeeded in destroying the last remnants of this once powerful community.

Scythian society. For administrative purposes the Royal Scyths divided their kingdom into four districts, each of which was controlled by a governor provided with a paid military bodyguard. The governor dispensed justice and levied taxes from the settlers and tribute from certain of the Greek Pontic cities. The tribal way of life remained in force; communities were governed by their elders or chieftains, these dignitaries being periodically summoned to attend a general assembly held in the presence either of the governor or of the sovereign. These assemblies differed from the gatherings held by the sovereign in the spring of each year in order to inspect and feast his army. In times of war the country was divided for purposes of recruitment into three sections, the enrolled men being later grouped into units, each one of which was commanded by its own officer.

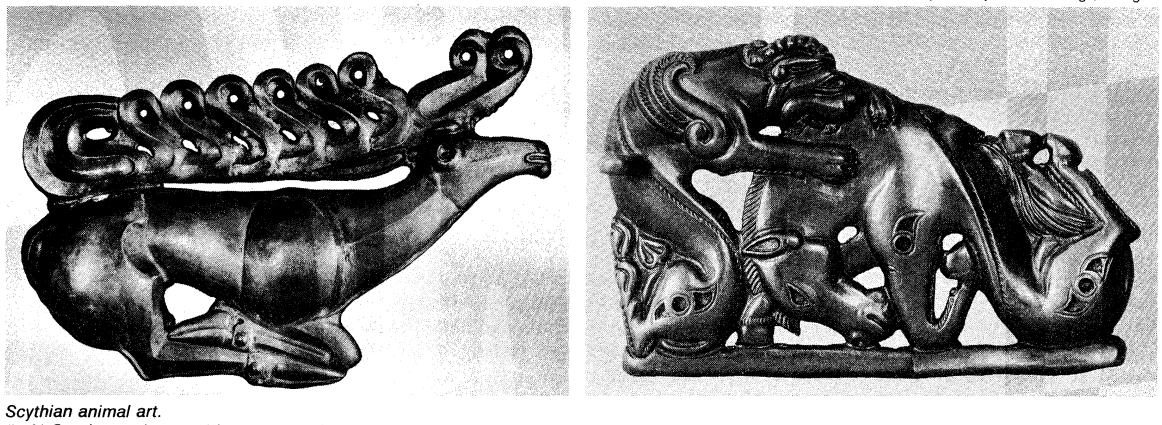
The army consisted entirely of freemen; *i.e.,* Scythian tribesmen; they were fed and clothed but paid no wage, though those of them who could produce the head of a soldier killed in battle were entitled to share in the day's booty. According to Herodotus, the Scythians scalped their victims, fashioning their skulls into cups that they wore attached to their belts, using them to pledge an oath in a mixture of blood and water. Such cups mounted in delicately worked gold have been found in several tombs.

Taking of scalps

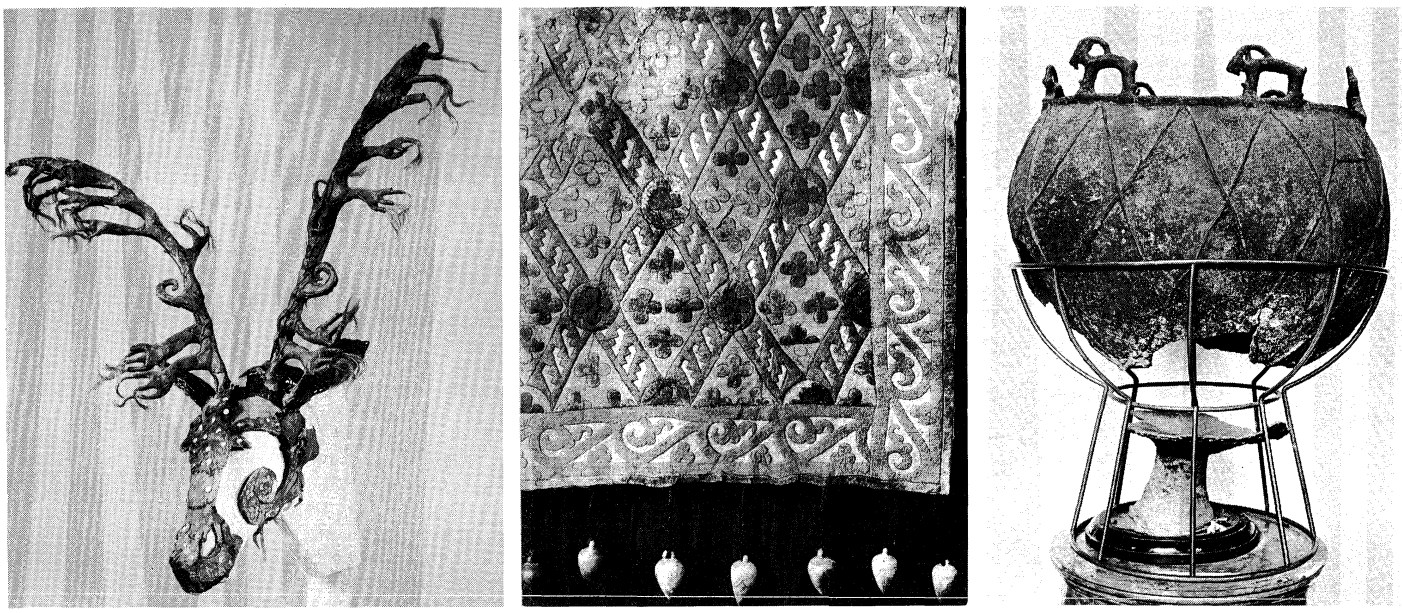
Many Royal Scyths wore bronze helmets and chain-mail jerkins of the Greek type lined with red felt. Their shields were generally round and made of leather, wood, or iron and were often decorated with a central gold ornament in the form of an animal, but other tribesmen carried square or rectangular ones. All used a double-curved bow, shooting over the horse's left shoulder; arrows had trefoil-shaped heads made, according to date, of bronze, iron, or bone. Arrows and bow were carried in a *gōrytos* (bow case) slung from the left side of the belt. Their swords were generally of the Persian type, with a heart-shaped or triangular crosspiece intricately ornamented. According to date, the blades were of bronze or iron; in southern Russia, the sheaths were often encased in gold worked into embossed designs and offset with paste or ivory inlay and gems. Their knives were of various shapes and lengths, some being curved in the Chinese manner. They wore the dagger attached to the left leg by straps, and many carried spears or standards surmounted by bronze terminals depicting real or imaginary beasts.

Every Scythian owned at least one gelding to serve as a riding horse, but the wealthy possessed a great many mounts; most Scythians also owned oxen or rough ponies, which served as beasts of burden. The finest riding horses were of the Fergana breed, but the majority were Mongolian ponies. The Scythians devoted much time and

By courtesy of the Hermitage, Leningrad



Scythian animal art.
(Left) Semirecumbent gold stag, probably central ornament from a shield, 6th century BC, from Kostromskaya Stanitsa; length 33.4 cm. (Right) Gold belt plate showing battle between two animals, 4th–3rd century BC, from Siberia; length 12.3 cm. In the Hermitage, Leningrad.



Artifacts preserved in the tombs of the Scythians.
(Left) Horse's burial mask of felt, leather, gilded hair, and copper in shape of stag's head, 5th century BC, from Pazyryk. (Centre) Section of felt appliqué wall hanging showing geometric pattern, 5th century BC, from Pazyryk. (Right) Bronze caldron, 4th century BC, from Chertomlyk; height 100 cm. In the Hermitage, Leningrad.
By courtesy of the Hermitage, Leningrad

attention to their horses and ornamented all their trappings. Bridles were provided with metal cheekpieces in the shape of animals, and the leather straps were adorned with embossed, cutout, or appliqué designs, which also often represented animals. Saddles consisted of two felt cushions mounted on wooden frames bound with yellow or red and sometimes embellished with gold plaques; the felt saddle cloths were adorned, as were the seats of the saddles, with appliqué designs. Metal stirrups were not known to the Scythians, but there is reason to believe that they rested their feet in felt or leather supports. Women travelled with their children in covered wagons with solid wheels and a central shaft along which mules or oxen could be yoked in pairs.

The frozen tombs of Pazyryk contained many well-preserved articles of clothing, all of which were profusely trimmed with complicated embroidered and appliqué designs; the clothes of the wealthy in southern Russia were covered with tiny gold-embossed plaques, sewn to the garments. At Pazyryk, felt appliqué wall hangings were found, some displaying religious scenes featuring the Great Goddess or anthropomorphic beasts, others with geometric or animal motifs. Even the embalmed body of a man was covered with tattooed designs of real and mythical beasts executed with great spirit and delicacy. Felt rugs were found, as well as a knotted, woollen pile carpet of Persian origin, from the 5th century BC, displaying figures of riders, elks, and stars. Felt cushions and mattresses; wooden tables with carved or turned legs and detachable, slightly hollowed, traylike tops; and wooden blocks serving as stools or head rests were customary.

The tombs of the Eurasian plain as a whole produced a mass of tools and domestic utensils, many of them made of precious materials. Rich jewelry and arms of great value and beauty are frequently found in the burials belonging to the western section. In addition, each burial throughout the entire area contained a cast bronze caldron of the distinctive Scythian shape. These caldrons vary in size from quite small examples to others weighing as much as 75 pounds; an overwhelming majority have a solid base, shaped like a truncated cone, around which the fire was heaped. The upper section is a semispherical bowl provided not with a loop handle but with handles (shaped like animals) fixed to the rim opposite each other. The finest caldron, found at Chertomlyk in the Dnepr (Dnieper) district, has six handles. At Pazyryk, small caldrons filled with stones and hempseeds were found standing beneath leather or felt tentlets with three or six supports. Herodotus referred to what he termed a Scythian purification rite that, he noted, consisted in inhaling the fumes of hempseeds thrown onto hot stones; the passage was well-nigh incomprehensible until archaeologists discovered that a smoking outfit of this sort had been provided for each person buried at Pazyryk, making it clear that hemp fumes were inhaled for pleasure and not, as Herodotus assumed, as part of a religious observance.

Hemp smoking

Occupations and beliefs. The Scythians were keen huntsmen and fishermen; they were skilled at curing hides; they excelled at working metals; and the settlers were good agriculturalists. Though they had neither an alphabet nor, until later times, a coinage, they carried on a lively trade not only with the inhabitants of Central Asia but also with the Greeks of the Pontic cities, often exchanging surplus goods and furs for Greek luxuries such as fine ceramic wares.

The Scythians worshipped the elements but they were not a devout people and never felt the need for temples. Their deepest feelings were centred on the Great Goddess, Tabiti-Hestia, the patroness of the fire and beasts, and she alone of all their deities figures in art.

Sarmatians. *Migration and conquests.* It is evident now that the Sarmatians advanced from central Asia to invade and conquer the Scythians and did not, as earlier thought, rise against them from within their own territory. It has become possible to divide Sarmatian history into four periods. The first extends from the 6th to the 4th century BC, during which the Sarmatians migrated from Asia and penetrated to the western foothills of the Urals. Toward the end of this period the Roxolani pushed on toward the Volga River and the Alani settled in the Kuban Valley. For the most part they remained till they were dislodged in the period of the great migration, when some of the latter infiltrated into Ossetia, surviving there as a community until about the 9th century AD.

Invasion of Scythia

The Sarmatians followed closely in the steps of these tribes, and by the 5th century BC they had already made themselves masters of the plain that stretches between the Urals and the Don River. This success encouraged them to intensify their pressure on the Scythians, and in the 4th century BC they crossed the Don and entered Scythia proper. This rapid advance may well have coincided with a decline in Scythian might, resulting from the death of their aged king, Ateas, in battle against Philip II of Macedon in 339 BC; if so, this would help to account for its speed.

The incursion serves to herald the second (Early Sarmatian) period in Sarmatian history, from the 4th to the 2nd century BC, often called the Prokhorov Period after the important burials of that name with which it is associated. During this period the Sarmatians partially overcame the Scythians and gradually succeeded them as rulers over most of southern Russia. The Roxolani had begun by joining with the Sarmatians in attacking the Scythians, advancing upon them from the south. Toward the end of this period, when the Sarmatians had strengthened their hold over the territory of the Royal Scyths, the Roxolani changed their allegiance, joining forces with the Scythians to attack the Greek Black Sea cities.

In the third (Middle Sarmatian) period, between the 1st century BC and the 1st century AD, the Sarmatians conquered all but the Crimean Scythians, bringing the assaults on the Greek cities to an end. Having consolidated their hold over the captured areas, the Sarmatians did not attempt to annex any of these cities; instead, early in Nero's reign, they invaded the Roman province of Lower Moesia (Bulgaria). Although the Roman general Plautius Silvanus Aelianus ejected them in AD 62–63, the Sarmatians and some Germanic tribes they had joined were thenceforth a menace to the Romans in the west, as were the Alani in the Caucasian area. Vespasian, Trajan, and Marcus Aurelius were forced to adopt such defensive measures as to attempt to transform the vassal Bosporan kingdom into a buffer state. Hadrian had to build a network of fortresses manned with Roman troops along the borders of Lower Moesia and Cappadocia.

In the last phase of their history, spanning the 2nd to 4th centuries AD, the Sarmatians and their Germanic allies entered Dacia (Romania) and began raiding the lower reaches of the Danube, but in the 3rd century the Gothic invasion put an end to their independence. Many Sarmatians nevertheless retained their position and influence under the Goths; others joined them to sweep into western Europe, fighting at their sides. Soon after AD 370, however, waves of migrating Huns effectively ended the very existence of Sarmatia; the majority of the Sarmatians who remained in southern Russia perished at the hands of these Asian invaders. Some of the survivors were assimilated by their new masters and others by the Slavs of the lower Dnepr, but most fled westward to join their neighbours in harassing the Huns and the remaining Goths. Their descendants continued to do so until the 6th century, when they finally disappeared from history.

Sarmatian society. The social structure or way of life developed by the Sarmatians at first followed closely that established by the Scythians, largely as the result of a common Central Asian origin and heritage. Nevertheless, several fundamental differences serve to distinguish the two peoples. Scythians venerated nature deities, while the Sarmatians were fire worshippers who sacrificed horses to their god. Scythian women were relegated to a life of semiseclusion; Sarmatian women, at least in the earlier periods, however, were expected to fight in time of war, and could not marry until they had killed an enemy in battle. After marriage, however, women were obliged to abandon warfare and to devote themselves entirely to their homes and families. Greek tales about Amazons may have been based on exploits of early Sarmatian women, many of whom were buried with their weapons.

Amazons

When the Sarmatians penetrated into southeastern Europe, they were already accomplished horsemen. They followed a nomadic way of life, devoting themselves to hunting and to pastoral occupations; most of them adhered to this form of existence to the end. There is evidence that some of them practiced an elementary form of husbandry, but few settlements of Sarmatian origin have as yet been studied. In the earlier periods their customs, clothes, and many other possessions resembled those of the Scythians, but their society was matriarchal in character. Gradually, as a prosperous class began to form, the tribe entered a transitional phase during which tribal chieftains tended to replace women as rulers; eventually, kings seem to have predominated.

The rise of a male monarchy may have been stimulated by the formation of a male corps of heavy cavalry, facilitated by the probably Sarmatian invention of a metal stirrup, soon followed by that of the spur. The methods of warfare thus made possible were largely responsible for Sarmatia's military supremacy. The Romans noted and gradually adopted certain of their tactics. During the 3rd century they went so far as to incorporate Sarmatian units in some Roman regiments, equipping them with their traditional weapons and encouraging them to fight in their own manner. From early times onward Sarmatian military accoutrements included conically shaped helmets worn with scale, ring, or plate armour of iron that often protected horses as well as riders. Unlike the Scythians, the Sarmatians did not excel at archery; they relied on long lances or spears and long, sharply pointed swords.

The Sarmatians were excellent craftsmen. Perhaps artistically less inventive than the Scythians, they were nevertheless equally proficient metalworkers, better potters, and no less adept at curing hides; they were thus able to maintain the important trade in furs, grain (levied from local settlers), honey, fish, and metal that the Scythians had established with the Greek cities on the northern shores of the Black Sea. The Sarmatians also developed commercial contact with the Syr-Darya region, the borderlands of China, and the kingdom of Khwārezm (Chorasmia).

Until disrupted by the Huns, Sarmatian culture presented a uniform and all-embracing character, even though it altered with each period in its history. Its evolution is reflected in burials that distinguish each of the four periods. Thus, no large mounds are associated with the earliest phase when objects were seldom included in the graves, though the occasional presence of articles points to the beginnings of a class society. Burials of the second period often display considerable wealth; those of the third phase also reflect the growth of a more complex way of life, stemming in part from the Sarmatian subjugation of alien tribes and assimilation of many of their customs. It also was a result of the flourishing trade the Sarmatians had established. During these periods the truly typical Sarmatian culture evolved and was imposed on the entire region. In the last phase Germanic influence led to the introduction of fibulae (brooches or clasps like safety pins), with resulting changes in costume.

Other Eastern peoples. These westward-moving peoples from the steppes are considered to have spoken Indo-European languages, but they may be regarded more as having been transitory overlords rather than substantial settlers. It is impossible to point to any one of them alone as exclusively ancestral to the first historical peoples of eastern Europe. The Slavs, whose historical documentation is later and less informative than that of the Celts and Teutonic peoples, probably represent the ultimate fusion of all the Indo-European elements on the western fringes of the steppeland. The people of the Lausitz culture may also have contributed, but the sequence of events remains very obscure. For the east Baltic Indo-European-speaking peoples, as represented by the Letts and Lithuanians, the archaeological evidence appears to indicate an undisturbed continuity of settlement going back to the time of the intrusion of Corded Ware people into that region.

(T.T.R./Ed.)

Greeks, Romans, and barbarians

The main treatment of classical Greek and Roman history is given in the article GRECO-ROMAN CIVILIZATION. Only a brief cultural overview is offered here, outlining the influence of Greeks and Romans on European history.

GREEKS

The history of Europe was determined by the diffusion throughout the continent of Indo-European tribes of European origin, of whom the Greeks were foremost as regards both the period at which they developed an advanced culture and their importance in further evolution. The Greeks came into being in the course of the 2nd millennium BC through the superimposition of a branch of the Indo-Europeans on the population of the Mediterranean region in the course of the great migrations of nations that started in the region of the lower Danube. From 1800 BC

onward the first early Greeks reached their later areas of settlement between the Ionian and the Aegean seas. The fusion of these earliest Greek-speaking people with their predecessors produced the civilization known as Mycenaean. They penetrated to the sea into the Aegean region and via Crete (approximately 1400 BC) reached Rhodes and even Cyprus and the shores of Asia Minor. From 1200 BC onward the Dorians followed from Epirus. They occupied principally parts of the Peloponnese (Sparta and Argolis) and also Crete. Their migration was followed by a dark age—two centuries of chaotic movements of tribes in Greece—at the end of which (*c.* 900 BC) the distribution of the Greek mainland among the various tribes was on the whole completed.

Mycenaean civilization

From *c.* 800 BC there was a further remarkable Greek expansion through the founding of colonies overseas. The coasts and islands of Asia Minor were occupied, from south to north, by the Dorians, the Ionians, and Aeolians respectively. In addition, individual colonies were strung out around the shores of the Black Sea in the north and across the eastern Mediterranean to Naukratis on the Nile Delta and in Cyrenaica and also in the western Mediterranean in Sicily, lower Italy, and Massilia (Marseille). Thus, the Hellenes, as they called themselves thereafter, came into contact on all sides with the old, advanced cultures of the Middle East and ultimately transmitted many features of these cultures to western Europe. By their own genius, however, the Greeks accomplished much more, and their achievements laid the foundations of European civilization.

The position and nature of the country exercised a decisive influence in the evolution of Greek civilization. The proximity of the sea tempted the Greeks to range far and wide exploring it, but the fact of their living on islands or on peninsulas or in valleys separated by mountains on the mainland confined the formation of states to small areas not easily accessible from other parts. This fateful individualism in political development was also a reflection of the Hellenic temperament. Though it prevented Greece from becoming a single unified nation that could rival the strength of the Middle Eastern monarchies, it led to the evolution of the city-state. This was not merely a complex social and economic structure and a centre for crafts and for trade with distant regions; above all it was a tightly knit, self-governing political and religious community, whose citizens were prepared to make any sacrifice to maintain their freedom. Colonies, too, started from individual cities and took the form of independent city-states. Fusions of power occurred in the shape of leagues of cities, such as the Peloponnesian League, the Delian League, and the Boeotian League. The efficacy of these leagues depended chiefly upon the hegemony of a leading city (Sparta, Athens, or Thebes), but the desire for self-determination of the others could never be permanently suppressed and the leagues broke up again and again.

Leagues of cities

The Hellenes, however, always felt themselves to be one people. In relation to the barbarians they were conscious of a common character and a common language, and they practiced only one religion. Furthermore, the great athletic contests and artistic competitions had a continually renewed unifying effect. The Hellenes possessed a keen intellect, capable of abstraction, and at the same time a supple imagination. They developed, in the form of the belief in the unity of body and soul, a serene, sensuous conception of the world. Their gods were only loosely connected by a theogony that took shape gradually; in the Greek religion there was neither revelation nor dogma to oppose the spirit of inquiry.

The Hellenes benefitted greatly from the knowledge and achievement of other countries as regards astronomy, chronology, and mathematics, but it was through their own native abilities that they made their greatest achievements, in becoming the founders of European philosophy and science. Their achievement in representative art and in architecture was no less fundamental. Their striving for an ideal, naturalistic rendering of the third dimension found its fulfillment in the representation of the human body in sculpture in the round. Another considerable achievement was the development of the pillared temple to a greater degree of harmony. In poetry the genius of the Hellenes created both form and content, which have remained a constant source of inspiration in European literature.

The strong political sense of the Greeks produced a variety of systems of government, from which their theory of political science abstracted types of constitution that are still in use. On the whole, political development in Greece followed a pattern: first, the rule of kings, found as early as the period of Mycenaean civilization; then a feudal period, the oligarchy of noble landowners; and finally, varying degrees of democracy. Frequently there were periods when individuals seized power in the cities and ruled as tyrants. The tendency for ever-wider sections of the community to participate in the life of the state brought into being the free democratic citizens, but the institution of slavery, upon which Greek society and the Greek economy rested, was untouched by this.

In spite of continual internal disputes the Greeks succeeded in warding off the threat of Asian despotism. The advance of the Persians into Europe failed (490 and 480–79 BC) because of the resistance of the Greeks and in particular of the Athenians. The 5th century BC saw the highest development of Greek civilization. The classical period of Athens and its great accomplishments left a lasting impression, but the political cleavages, particularly the struggle between Athens and Sparta, increasingly reduced the political strength of the Greeks. Not until they were conquered by the Macedonians did the Greeks attain a new importance as the cultural leaven of the Hellenistic empires of Alexander the Great and his successors. A new system of colonization spread as far as the Indus city-communities fashioned after the Greek prototype, and Greek education and language came to be of consequence in the world at large.

Greece again asserted its independence through the formation of the Achaean League, which was finally defeated by the Romans in 146 BC. The spirit of Greek civilization subsequently exercised a great influence upon Rome. Greek culture became one of the principal components of Roman imperial culture, and together with it spread throughout Europe. When Christian teaching appeared in the Middle East, the Greek world of ideas exercised a decisive influence upon its spiritual evolution. From the time of the partition of the Roman Empire leadership in the Eastern Empire fell to the Greeks. Their language became the language of the state, and its usage spread to the Balkans. The Byzantine Empire, of which Greece was the core, protected Europe against potential invaders from Asia Minor until the fall of Constantinople in 1453.

ROMANS

The original Mediterranean population of Italy was completely altered by repeated superimpositions of peoples of Indo-European stock. The first Indo-European migrants, who belonged to the Italic tribes, moved across the eastern Alpine passes into the plain of the Po River about 1800 BC. Later they crossed the Apennines and eventually occupied the region of Latium, which included Rome. Before 1000 BC there followed related tribes, which later divided into various groups and gradually moved to central and southern Italy. In Tuscany they were repulsed by the Etruscans, who may have come originally from Asia Minor. The next to arrive were Illyrians from the Balkans, who occupied Venetia and Apulia. At the beginning of the historical period Greek colonists arrived in Italy, and after 400 BC the Celts, who settled in the plain of the Po.

The city of Rome, increasing gradually in power and influence, created through political rule and the spread of the Latin language something like a nation out of this abundance of nationalities. In this the Romans were favoured by their kinship with the other Italic tribes. The Roman and Italic elements in Italy, moreover, were reinforced in the beginning through the founding of colonies by Rome and by other towns in Latium. The Italic element in Roman towns decreased: a process—less racial than cultural—called the Romanization of the provinces. In the 3rd century BC central and southern Italy were dotted with Roman colonies, and the system was to be extended to ever more distant regions up to imperial times.

Romanization of the provinces

As its dominion spread throughout Italy and covered the entire Mediterranean basin, Rome received an influx of people of the most varied origins, including eventually vast numbers from Asia and Africa.

The building of an enormous empire was Rome's greatest achievement. Held together by the military power of one city, in the 2nd century AD the Roman Empire extended throughout northern Africa and western Asia; in Europe it covered all the Mediterranean countries, Spain, Gaul, and southern Britain. This vast region, united under a single authority and a single political and social organization, enjoyed a long period of peaceful development. In Asia, on a narrow front, it bordered the Parthian Empire, but elsewhere beyond its perimeter there were only barbarians. Rome brought to the conquered parts of Europe the civilization the Greeks had begun, to which it added its own important contributions in the form of state organization, military institutions, and law. Within the framework of the empire and under the protection of its chain of fortifications, extending, uninterrupted, the entire length of its frontiers (marked in Europe by the Rhine and the Danube), there began the assimilation of varying types of culture to the Hellenistic-Roman pattern. The army principally, but also Roman administration, the social order, and economic factors, encouraged Romanization. Except around the eastern Mediterranean, where Greek remained dominant, Latin became everywhere the language of commerce and eventually almost the universal language.

The empire formed an interconnected area of free trade, which was afforded a thriving existence by the *pax romana* ("Roman peace"). Products of rural districts found a market throughout the whole empire, and the advanced technical skills of the central region of the Mediterranean spread outward into the provinces. The most decisive step toward Romanization was the extension of the city system into these provinces. Rural and tribal institutions were replaced by the *civitas* form of government, according to which the elected city authority shared in the administration of the surrounding country region; and, as the old idea of the Greek city-state gained ground, a measure of local autonomy appeared. The Romanized upper classes of the provinces began supplying men to fill the higher offices of the state. Ever-larger numbers of people acquired the status of Roman citizens, until in AD 212 the emperor Caracalla bestowed it on all freeborn subjects. The institution of slavery, however, remained.

The enjoyment of equal rights by all Roman citizens did not last. The coercive measures by which alone the state could maintain itself divided the population anew into hereditary classes according to their work; and the barbarians, mainly Germanic, who were admitted into the empire in greater numbers, remained in their own tribal associations either as subjects or as allies. The state created a perfected administrative apparatus, which exercised a strongly unifying effect throughout the empire, but local self-government became less and less effective under pressure from the central authority.

The decline of the late empire was accompanied by a stagnation of spiritual forces, a paralysis of creative power, and a retrograde development in the economy. Much of the empire's work of civilization was lost in internal and external wars. Equally, barbarization began with the rise of unchecked pagan ways of life and the settlement of Germanic tribes long before the latter shattered the Western Empire and took possession of its parts. Though many features of Roman civilization disappeared, others survived in the customs of peoples in various parts of the empire. Moreover, something of the superstructure of the empire was taken over by the Germanic states, and much valuable literature was preserved in manuscript for later times.

The spread of Christianity

It was under the Roman Empire that the Christian religion penetrated into Europe. By winning recognition as the religion of the state, it added a new basic factor of equality and unification to the imperial civilization and at the same time reintroduced Middle Eastern and Hellenistic elements into the West. Organized within the framework of the empire, the church became a complementary body upholding the state. Moreover, during the period of the decline of secular culture, Christianity and the church were the sole forces to arouse fresh creative strength, by assimilating the civilization of the ancient world and transmitting it to the Middle Ages. At the same time the church in the West showed reserve toward the speculative dogma of the Middle Eastern and Hellenic worlds and directed its attention more toward questions of morality and order. When the Western Empire collapsed and the use of Greek had died there, the division between East and West became still sharper. The name Romaioi remained attached to the Greeks of the Eastern Empire, while in the West the word Roman developed a new meaning in connection with the church and the bishop of Rome. Christianity and a church of a Roman character, the most enduring legacy of the ancient world, became one of the most important features in western European civilization.

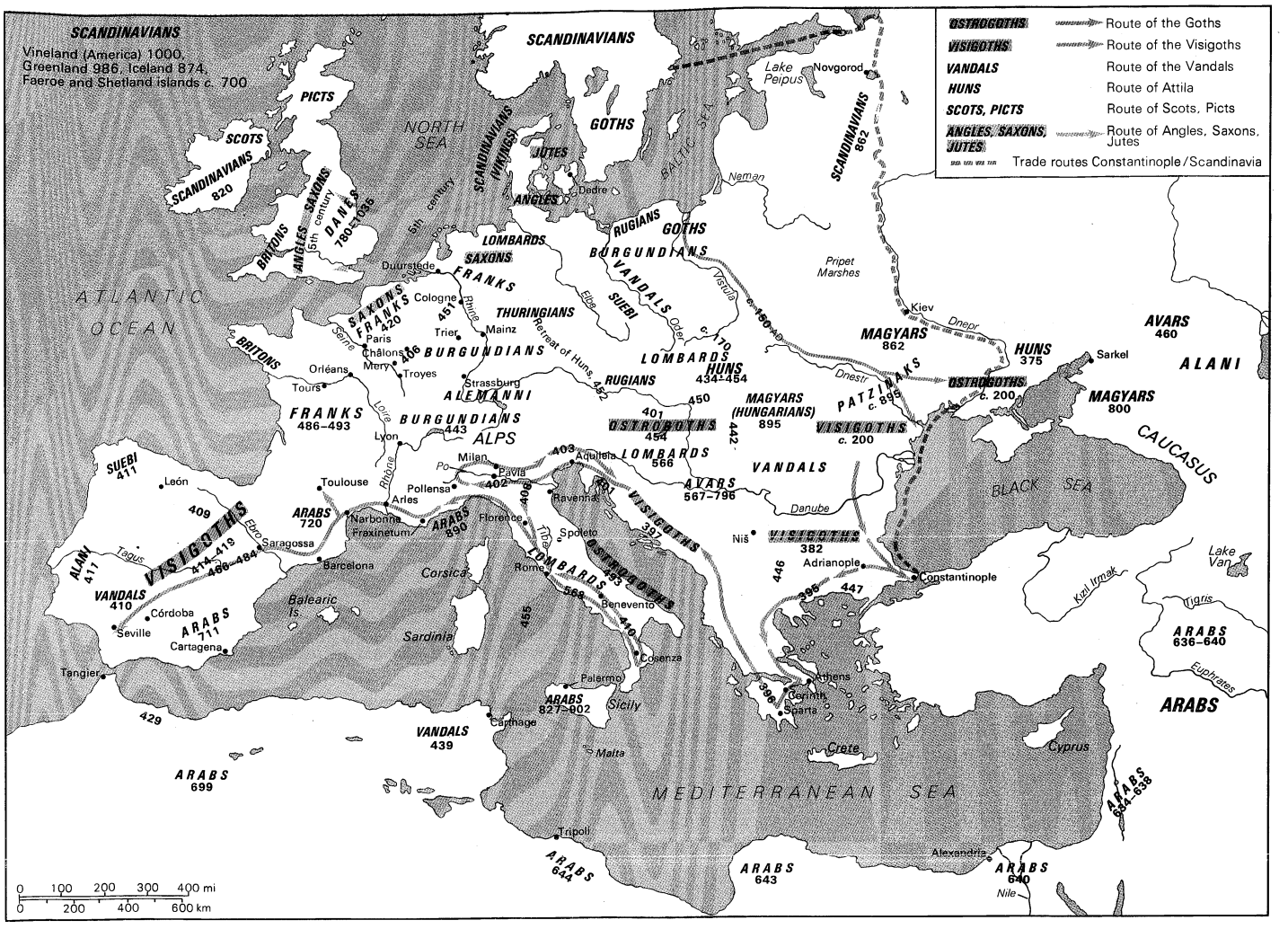
BARBARIAN MIGRATIONS AND INVASIONS

The Germans and Huns. The wanderings of the Germanic peoples, which lasted until the early Middle Ages and destroyed the Western Roman Empire, were, together with the migrations of the Slavs, formative elements of the distribution of peoples in modern Europe. The Germanic peoples originated about 1800 BC from the superimposition, on a population of megalithic culture on the eastern North Sea coast, of Battle-Ax people from the Corded Ware culture of middle Germany. During the Bronze Age the Germanic peoples spread over southern Scandinavia and penetrated more deeply into Germany between the Weser and Vistula rivers. Contact with the Mediterranean through the amber trade encouraged the development from a purely peasant culture, but during the Iron Age the Germanic peoples were at first cut off from the Mediterranean by the Celts and Illyrians. Their culture declined and an increasing population together with worsening climatic conditions drove them to seek new lands farther south. Thus the central European Celts and Illyrians found themselves under a growing pressure. Even before 200 BC the first Germanic tribes had reached the lower Danube, where their path was barred by the Macedonian kingdom. Driven by rising floodwaters, at the end of the 2nd century BC migratory hordes of Cimbri, Teutoni, and Ambrones from Jutland broke through the Celtic-Illyrian zone and reached the edge of the Roman sphere of influence, appearing first in Carinthia (113 BC), then in southern France, and finally in upper Italy. With the violent attacks of the Cimbri the Germans stepped onto the stage of history.

The nature of the Germanic migrations

These migrations were in no way nomadic; they were the gradual expansions of a land-hungry peasantry. Tribes did not always migrate en masse. Usually, because of the loose political structure, groups remained in the original homelands or settled down at points along the migration route. In course of time, many tribes were depleted and scattered. On the other hand, different tribal groups would sometimes unite before migrating or would take up other wanderers en route. The migrations required skilled leadership, and this promoted the social and political elevation of a noble and kingly class. In 102 BC the Teutoni were totally defeated by the Romans, who in the following year destroyed the army of the Cimbri. The Swabian tribes, however, moved steadily through central and southern Germany, and the Celts were compelled to retreat to Gaul. When the Germans under Ariovistus crossed the upper Rhine, Julius Caesar arrested their advance and initiated the Roman countermovement with his victory in the Sundgau (58 BC). Under the emperor Augustus, Roman rule was carried as far as the Rhine and the Danube. On the far side of these rivers the Germans were pushed back only in the small area contained within the Germano-Raetian limes (fortified frontier) from about AD 70.

The pressure of population was soon evident once more among the German peoples. Tribes that had left Scandinavia earlier (Rugii, Goths, Gepidae, Vandals, Burgundians, and others) pressed on from the lower Vistula and Oder rivers (AD 150 onward). The unrest spread to other tribes, and the resulting wars between the Romans and the Marcomanni (166–180) threatened Italy itself. The successful campaigns of Marcus Aurelius resulted in the ac-



Migrations and conquests, 2nd to 10th century.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

quisition by Rome of the provinces of Marcomannia and Sarmatia, but after his death these had to be abandoned and the movement of the Germanic peoples continued. Soon the Alamanni, pushing up the Main River, reached the upper German limes.

To the east the Goths had reached the Black Sea about AD 200. Year after year Goths and others, either crossing the lower Danube or travelling by sea, penetrated into the Balkan Peninsula and Asia Minor as far as Cyprus on plundering expeditions. Only with the Roman victory at Nish (269) was their advance finally checked. Enriched with booty and constituted imperial mercenaries in return for a yearly tribute, they became a settled population. The Romans, however, surrendered Dacia beyond the Danube.

In 258 the Alamanni and the Franks broke through the lines and settled on the right bank of the Rhine, continuously infiltrating thereafter toward Gaul and Italy. Everywhere within the empire towns were fortified, even Rome itself. Franks and Saxons ravaged the coasts of northern Gaul and Britain, and for the next three centuries incursions by Germanic peoples were the scourge of the Western Empire. Nevertheless, it was only with German help that the empire was able to survive as long as it did. The Roman army received an ever-growing number of recruits from the German tribes, which also provided settlers for the land. The Germans soon proved themselves capable of holding the highest ranks in the army. Tribute money to the tribes, pay to individual soldiers, and booty all brought wealth to the Germans, which in turn gave warrior lords the means with which to maintain large followings of retainers. In the West, however, among the Alamanni and Franks, the beginnings of political union into larger groups did not go beyond loose associations. Only in the East did the Gothic kingdom gather many tribes under a single leadership. Above all, the development of the East Germans was stimulated by their undisturbed contact with the frontiers of the ancient world. Their economy, however, was still unable to support a steadily growing population, and pressure from overpopulation resulted in further incursions into the Roman Empire. The imperial reforms of Diocletian and Constantine the Great brought a period of improvement. The usurpation of the imperial title by a Frankish general in 356 let loose a storm along the whole Rhine and subsequently on the Danube, but the frontiers were restored by the emperors Julian and Valentinian I, who repelled attacks by both the Franks and the Alamanni.

The advent of the Huns

Now, however, a new force appeared. In 375 the Huns from Central Asia first attacked the Ostrogoths—an event that provoked serious disturbances among the East Germans. The Huns remained in the background, gradually subjugating many Germanic and other tribes. The terrified Goths and related tribes burst through the Danube frontier into the Roman Empire, and the Balkans became once again a battlefield for German armies. After the crushing defeat of the Romans at Adrianople (378), the empire was no longer in a position to drive all its enemies from its territories. Tribes that could no longer be expelled were settled within the empire as "allies" (*foederati*). They received subsidies and in return supplied troops. The Germanization of the empire progressed, that of the army being nearly completed. None of the tribes, however, that had broken into the Balkans settled there. After the division of the empire in 395, the emperors at Constantinople did all in their power to drive the Ger-

manic tribes away from the vicinity of the capital toward the Western Empire.

From the beginning of the 5th century the Western Empire was the scene of numerous further migrations. The Visigoths broke out of the Balkans into Italy and in 410 temporarily occupied Rome. In 406–407 Germanic and other tribes (Vandals, Alani, Suebi, and Burgundians) from Silesia and even farther east crossed the Rhine in their flight from the Huns and penetrated as far as Spain. The Vandals subsequently crossed to Africa and set up at Carthage the first independent German state on Roman soil. In the Battle of the Catalaunian Plains (451) the Roman commander Aëtius, with German support, defeated Attila, who had united his Huns with some other Germans in a vigorous westward push. The Balkans suffered a third period of terrible raids from the East Germans; and Jutes, Angles, and Saxons from the Jutland Peninsula crossed over to Britain. The Franks and the Alamanni finally established themselves on the far side of the Rhine, the Burgundians extended along the Rhône Valley, and the Visigoths took possession of nearly all Spain. In 476 the Germanic soldiery proclaimed Odoacer, a barbarian general, as king of Italy, and when Odoacer deposed the emperor Romulus Augustulus at Ravenna, the empire in the West was at an end.

The restoration of the imperial power in the West by Justinian I, who reconquered Africa, Italy, and parts of Spain, was of short duration. The Visigoths regained their position in Spain; Africa was lost to the Arabs; and in 568 the Lombards conquered upper and central Italy, Rome and Ravenna alone excepted. The Lombards had been driven from the plains of the Danube and Tisa by a new wave of mounted Asian nomads, the Avars. The Germanic migrations on the European mainland were finally at an end. All the Roman territories between the Danube and Rhine and the Atlantic had been penetrated by the Germans. In western Europe a new ethnic pattern was established, which subsequently was scarcely altered.

Three zones of Germanic influence

The Germanic tribes were roughly distributed into three zones. The central part of Germany was composed of peoples of purely Germanic stock. In the second zone a mixture of Germanic peoples ruled in Britain, in the country west of the Rhine as far as the Seine basin, and in the area south of the Danube toward the Alps. In the third zone (southern France, Italy, and Spain) the German element was much dispersed among the Romance peoples.

The migrations had meant that at first only the Germans' warlike qualities were revealed, and their manifold talents only gradually became apparent. But originating with the Goths, a characteristic decorative art developed which, though eclectic, gave expression to the spirit of the Germanic ruling class. Together with an epic poetry, purely Germanic in origin, it reached its highest point with the growth of independent Germanic states, and its influence was carried on into the Middle Ages. The Germanic states, with their simpler political systems, saved Western society from authoritarian oppression. Beginning with the great translation of the Bible by Ulfilas, the Goths created a religious literature, which, however, was soon destroyed because of its Arianism. Nevertheless, codes of law that, with the help of Roman jurists and the Latin language, were promulgated by the Visigoths (*c.* 450), by the Ostrogoths (508), and by the Franks (*lex salica,* 508–511) survived. To a certain extent the Germans were able to take over constituent parts of the Roman state. Through their conversion to Christianity they became members of the Roman Church, which had most fully inherited the culture of antiquity.

The Slavs, Bulgarians, and Hungarians. The Slavs were the last of the Indo-Europeans to settle in their present home. They originally occupied the area between the Vistula and Dnepr rivers, stretching northward from the Carpathian Mountains as far as the Narew River. The Slavs were first mentioned by name in 518, when they broke into the Roman Empire by way of the lower Danube. Their social structure was simple, consisting of small tribal units. Their further expansion was caused partly by the Avars, who from 568 were the sole rulers of the Hungarian plain, holding the surrounding Slavs in subjection. In 623–624 the Slavs overthrew the Avar lordship in Bohemia. Gradually they established frontiers with the German tribes in the Eastern Alps at the Bohemian Forest and along the Saale and Elbe rivers. By the end of the 8th century at the latest, Slav settlements stretched from the east coast of Holstein as far as the Vistula. The weakness of the Byzantine Empire made possible further attacks across the Danube. Even after the heavy defeat of the Avars before Constantinople in 626, the Slavs remained in the Balkan Peninsula under Byzantine lordship. The Balkans became a Slav territory, the Greeks fell back on the eastern and southern coasts, and the Romans fell back on the Adriatic or remained in isolation as "Balkan Romanians."

In 679 the Bulgarians, or Bulgars, a Turanian people from southern Russia, burst across the lower Danube and conquered the eastern part of the Balkan Peninsula. They were quickly assimilated to the Slavs, however, and with them were converted to Christianity by the end of the 9th century. A new invasion from the 9th century onward of Asian nomads, the Hungarians, or Magyars, divided the Slavs of the south from those of the west and the east. The West Slavs, moreover (Czechs, Slovaks, Elbe Slavs, Poles, and Pomeranians), received Christianity from the west, and this separated them from the East Slavs (White Russians, Ukrainians, and Great Russians), who adhered to the Greek Orthodox Church. After the Mongol invasions from 1223 onward the East Slavs withdrew from the Dnepr territories toward the wooded lands to the northwest, where they drove back the Finns and themselves developed as the Great Russians. (H.Au.)

The Middle Ages

The ancient world had unified around the Mediterranean—politically by the supremacy of Rome and culturally by a common Greco-Roman or Hellenistic literature, thought, and art. This unity was never complete, and signs of disintegration were clearly apparent during the 4th century; within another 250 years the unity of the Mediterranean was permanently broken. Thus antiquity was both more and less than the beginning of Western civilization—more because it was also the beginning of the Byzantine and the Islāmic civilizations, less because Western civilization included elements that were lacking or not fully assimilated in the Roman Empire at its height. Medieval Europe expanded far beyond the geographical limits of antiquity and included peoples who were never under Roman rule, and its culture was dominated by Christianity, which originated within the empire as an element alien to pagan and classical antiquity. The heritage of Rome was fused with the Germanic and Christian elements to provide the medieval foundations of Western civilization.

THE CHRONOLOGY OF THE MIDDLE AGES

Germanic kingdoms. The decline of Roman authority in the West, culminating in the deposition of Romulus Augustulus (476) and the establishment of Germanic kingdoms in its place is described above. The eastern provinces of the empire fared better during the period of invasions after 378: the emperors in Constantinople ensured their safety not only by arms but also by a diplomacy aimed at sending troublesome tribes westward to gain their fortunes. Thus in 489 Zeno granted the title of patrician to Theodoric, king of the Ostrogoths, and gave him permission to conquer Italy and rule there in the emperor's name.

The Ostrogoths under Theodoric

The kingdom of the Ostrogoths under Theodoric (493–526) was the most advanced of the barbarian kingdoms. Like the others, but far more successfully, the Ostrogothic kingdom strove to preserve as much of Roman institutions and culture as its more limited resources allowed.

The last great Germanic conquest and the most lasting was that of Clovis, who led his Franks into northern Gaul and by 507 had pushed the Visigoths south into Spain. Only the kingdom of Burgundy and the Mediterranean coastal strip (Septimania and Provence) lay outside the Frankish Merovingian kingdom. Perhaps the most important reason for Clovis' success, apart from his ability as a warrior, was his policy toward the church. Unlike

most of the other Germanic tribes, who were Christians of the heretical Arian sect, the Franks were still heathen when they crossed into Gaul. The conversion of Clovis to the Catholic faith was a decisive moment in the history of western Europe. He was now the only Catholic ruler among the Germanic kings and thus gained what no other ruler could: the support of the church and the Gallo-Roman subject population. His conquest of Gaul became a war of liberation from the yoke of the hated Arian heretics.

In the 6th century none of the Germanic kingdoms achieved political stability, except for brief periods under exceptionally strong rulers. In Gaul the Merovingian kingdom was divided among the sons after the death of the king, a provision that almost guaranteed continual war for supremacy. In Anglo-Saxon England, where partible inheritance was not the rule, there were yet so many original petty kingdoms (Bernicia, Deira, Lindsey, Mercia, East Anglia, Essex, Wessex, Kent, and several other regions whose tribal leaders briefly claimed royal status) that warfare there was also endemic, as each ruler strove to maintain and expand the political boundaries of his kingdom. On the continent, except in Merovingian Gaul, the barbarian kingdoms declined. Burgundy was conquered by the Merovingians; the Byzantine emperor Justinian destroyed the Vandal kingdom, reconquered Ostrogothic Italy, and temporarily reoccupied the southern part of Visigothic Spain in campaigns lasting from 533 to 556. But Italy was so ravaged that the real beneficiaries of Justinian's policy were the Lombards, the last of the Germanic peoples to settle within the limits of the Roman Empire. Beginning in 568, the Lombard conquest was rapid, and by about 600 only a few districts, such as Ravenna, Sicily, and Rome, remained nominally subject to the Byzantine Empire, organized as the exarchate of Ravenna.

Except in England, where the vestiges of late Roman civilization were almost wiped out, the barbarian conquerors did not purposefully destroy the Roman system that they took over; rather, they were unable to maintain it. The *civitas* (city with surrounding countryside) continued to be the unit of local government, but the local count had only loose ties to the central royal government, and it was so difficult to collect the old Roman taxes that they eventually disappeared. In the stagnant economy, however, this was not serious, for the barbarian government performed almost no service other than to maintain some semblance of law and order. The law was twofold, in accordance with the Germanic concept of "personality of law." The provincial Roman population lived under codes derived from the already simplified version of Roman law in the Theodosian Code (*c.* 438), while the various Germanic tribes lived under their *leges barbarorum* ("laws of the barbarians"). The latter, written in Latin, were also influenced by Roman-law concepts.

Genesis of Latin Christendom. Christianity originated as an offshoot from Judaism, in an eastern province of the Roman Empire. It began as a religion profoundly hostile to many of the values and vested interests of the ancient world. At first merely suspect and despised, later actively repressed by the Roman government, within three centuries it had spread throughout the empire, and the emperor was a Christian. Christianity had absorbed many of the best cultural and institutional elements of antiquity, which it was the historic role of the church to transmit to the Middle Ages. By the end of the 4th century, not only was Christianity the official religion of the empire but all other religions were proscribed.

The gradual division of Christianity into a Latin and a Greek communion was part of the breakdown of the cultural unity of the Mediterranean. By the end of the 3rd century there were few educated men who could read and speak both Greek and Latin. Equally important in the growth of a distinctly Latin church was the rise of the papacy between the 4th and 7th centuries. The popes were recognized as specially preeminent in the West by the emperors of the 4th century, and papal supremacy was recognized by the Council of Constantinople (381), though such recognition was not accepted in the Greek East. The absence of the imperial government, now in Constantinople, and the consistent record of the bishops of Rome for conservatism and orthodoxy during the early heresies also contributed to the rise of the papacy. At an early date the Petrine theory of the supremacy of the pope as St. Peter's successor was widely accepted, though the first pope to enunciate the doctrine formally was Leo I the Great, in the 5th century.

The rise of the papacy

With the decline of imperial authority in Italy the pope became practically an independent secular ruler of Rome, and with the disintegration of the empire into barbarian kingdoms, the church, with the pope at its head, became the most important element of unity in the west. Latin Christendom grew with the conversion of the barbarians to the Catholic faith (either from Arianism or from heathen religious beliefs). Conversion was actively promoted by missionaries of the monastic order founded by St. Benedict, about 520, and their work was often sponsored by the papacy. Especially notable was the conversion of the Anglo-Saxons in the 7th century and of the Germans in the 8th. In each case the papacy either initiated or supported a movement undertaken by monks.

The art, literature, and thought of the early Middle Ages were almost completely dominated by the interests of the church. The age of the Latin Fathers, whose greatest achievement in theology was reached in the works of Ambrose, Jerome, and Augustine, was followed by a long cultural decline. The three centuries from about 450 to around 750 were the Dark Ages of western Europe, but even in this period not all was darkness. The monks not only spread the faith but also, in their monasteries, preserved the literary remains of classical antiquity as well as the writings of the Fathers. Among the writers who had the greatest influence upon or interest for the future of medieval culture were Boëthius, Gregory of Tours, Pope Gregory the Great, Isidore of Seville, and the best scholar of the whole period from the decline of Rome to the high Middle Ages, the Northumbrian historian and theologian, Bede.

Final breakdown of Mediterranean unity. The Greek half of the Roman Empire did not suffer, relatively, much economic or political loss from the barbarian invasions that inaugurated the early Middle Ages in the West. In fact, the Byzantine Empire preserved much of the civilization of antiquity and added elements from the Hellenistic and Oriental world that further differentiate the Byzantine East from the Latin West (see BYZANTINE EMPIRE). Under Justinian I (527–565) the Byzantine Empire enjoyed a prosperity and a flourishing of the arts that made later generations think of that period as a golden age. But Justinian failed in his greatest ambition, to reconquer the lost western provinces and reunite the Mediterranean. His greatest monuments were not military and political, as he hoped, but cultural: the great church of Hagia Sophia and the codification of the Roman law in the *Corpus juris civilis.* Actually, the Byzantine Empire had weaknesses that neither Justinian nor his successors were able to overcome. Most notable was the bitter and sometimes violent religious dissension between the Greek and Orthodox people of Constantinople and the nearby provinces and the various Christian sects in the provinces of Egypt, Syria, and parts of Asia Minor. Also, for centuries, danger from Persia had drained Byzantine resources; finally, Heraclius (610–641) was triumphant in a great war against Persia, but the empire emerged almost wholly exhausted.

The rise of Islām

In the struggle between Persia and the Byzantine Empire the ultimate victor was Islām. The prophet Muḥammad in the 7th century united the Arabs under his religious and military leadership. In less than a hundred years his successors, the caliphs, conquered lands that stretched all the way from southern Gaul and Spain to western India. The Byzantine provinces of North Africa, Egypt, Syria, and Palestine (where religious antagonism toward Constantinople undermined resistance against the relatively tolerant Arabs) fell quickly; all of Persia was overrun in a few years. Christendom—both Greek Orthodox and Latin—was confronted with a successful and hostile Arab empire whose unity and strength derived from a common religion, Islām, and a common Arab culture and language.

The breakdown of Mediterranean unity was completed

by the middle of the 8th century. The Merovingian kings had degenerated in *rois fainéants,* and their power had passed to the Carolingian mayors of the palace. The popes, threatened by the Lombards who sought to conquer all territories still under Byzantine sovereignty, appealed in vain to Constantinople for help. The Byzantine emperors, hard pressed to defend their frontiers against the Muslims, could spare no troops to rescue Rome from Lombard conquest. Furthermore, the emperor Leo III the Isaurian (717–741) inaugurated a religious policy, iconoclasm, that Latin Christians considered heretical; Leo and his successors attempted to force iconoclasm on the pope. Confronted with this situation, the pope sought help from another quarter—the Carolingian mayor in Merovingian Gaul, Pepin III the Short, who was eager to win the pope's support for his desire to supplant the Merovingians and become king of the Franks. The alliance between papacy and Frankish monarchy was completed when Pepin was consecrated king in return for intervention in Italy to save Rome from the Lombards. By transferring to the pope the former Byzantine territory that he reconquered from the Lombards, Pepin established the Papal States under the pope as temporal ruler. The pope then conferred upon Pepin the title of *patricius Romanorum.* To justify these transactions (which, to be legal, should have been accomplished or at least approved by the Byzantine emperor), the papal chancery or someone familiar with its practice produced the Donation of Constantine, the most famous forgery of all time. It purported to be a grant of supreme authority in the West from the emperor Constantine to Pope Sylvester I and his successors.

Carolingian Empire. As king, Pepin extended his power over all of Gaul by reducing counts and dukes to obedience and pushing the Muslims back across the Pyrenees. His successor was Charlemagne (Charles I), who subdued all the hitherto independent German tribes, reduced the Slavonic peoples on the eastern border to a tributary status, conquered the Lombard kingdom in Italy (taking the title of king of the Lombards for himself), and established a Frankish county south of the Pyrenees in Spain. His kingdom now stretched from the Elbe-Saale line in the northeast to the Ebro in the southwest and from the North Sea to the marches of Benevento in southern Italy. Except in the British Isles, there was no territory inhabited by Christians in western Europe that did not owe allegiance to Charlemagne. He was acclaimed the "leader of the Christian people" and "defender of Christ's churches." Except for purely spiritual matters, Charlemagne considered himself ruler of both church and state. He used bishops and abbots as officials of the government; and his laws, the capitularies, regulated the affairs of both clergy and laity. His officials and advisers were already beginning to refer to his dominions as a "Christian empire" and to him as successor of Constantine when in the year 800 he went to Rome to put down a local uprising against Pope Leo III. On Christmas day, unexpectedly, Leo crowned him, and the assembled clergy and people acclaimed him Augustus and emperor. The actual power and nature of his authority were not changed by the act, which merely recognized and formalized the position he had attained.

Cultural revival under Charlemagne

The revival of cultural activities under Charlemagne and his successors was so impressive that the period from about 775 to about 875 has been called a renaissance. Charlemagne's immediate purpose was to provide a better educated clergy, as part of his concern for the general welfare of the church. He encouraged monasteries and cathedrals to maintain schools, and he made his court a centre of learning by inviting scholars and writers from Ireland, England, Spain, and Italy, as well as Gaul. The most important of these men was Alcuin of York, who later became abbot of St. Martin's of Tours, where he reformed the degenerated handwriting of the times. The new style, called Carolingian minuscule, was employed for the copying of most of the manuscripts of the classical works that have survived. It was also in this period that the seven liberal arts were organized into the formal curricula that remained until the 16th century. These were the *trivium* (grammar, rhetoric, and dialectic) and the *quadrivium* (arithmetic, geometry, astronomy, and music). Little that was original or profound was produced in the Carolingian Renaissance; in the cultural history of Europe it marks a stage in which preservation of the heritage of both pagan and Christian antiquity was the main achievement.

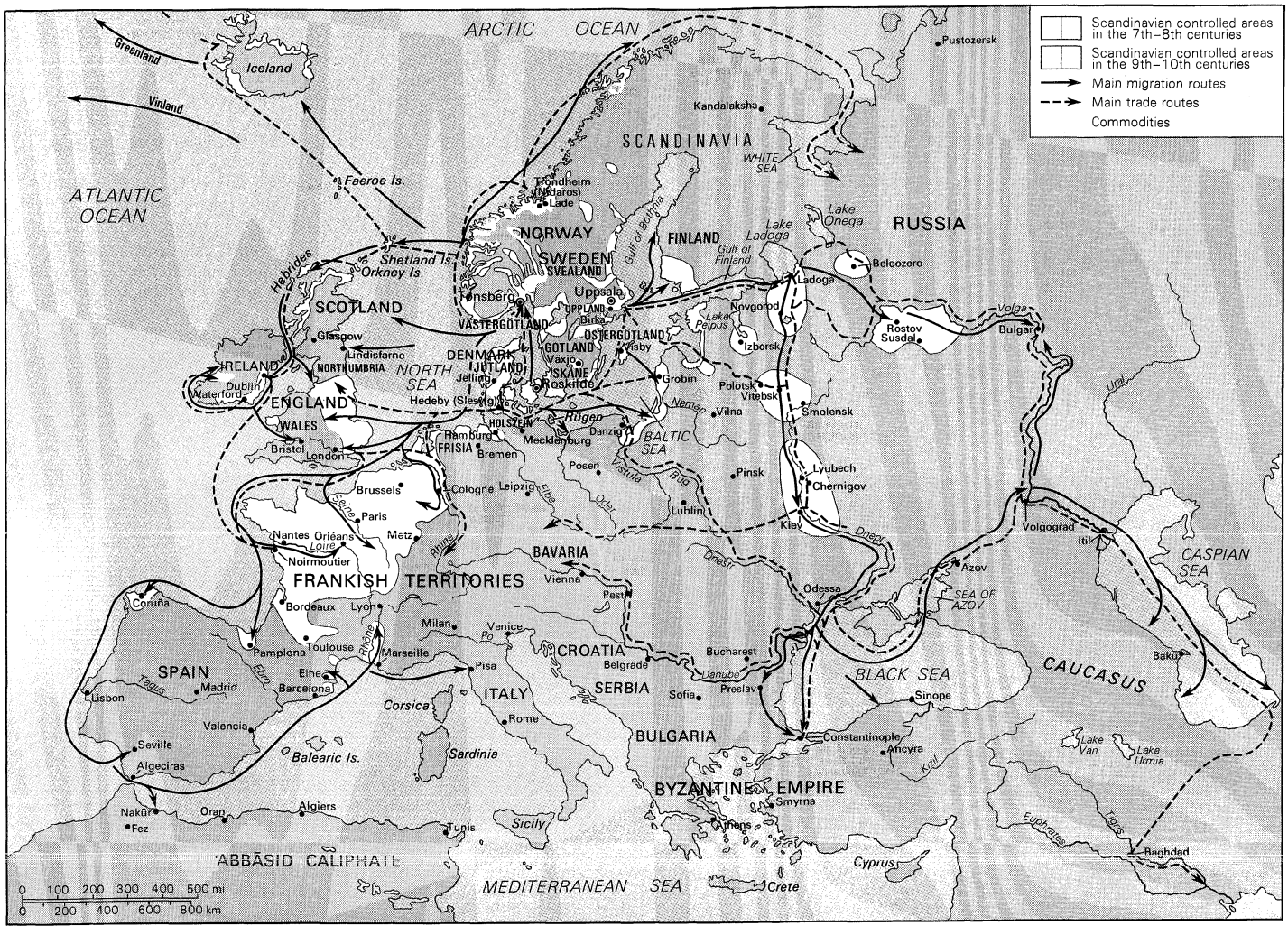
Collapse of the Carolingian Empire. Imposing as it was, Charlemagne's great empire had real weaknesses: geographically it was too large for the rudimentary governmental system; economically it was backward; in communications, transport, and military organization it was inferior to the Roman Empire. Also, among the nobility there were deep-rooted desires to control local affairs independently of the central government and the rest of the empire. A fatal flaw in the system was that it needed a Charlemagne to make it work. These were the underlying causes of the disintegration of the empire after Charlemagne's death. The direct causes were civil wars among his successors and a new wave of barbarian invasions.

The unity of the empire was formally preserved when Charlemagne was succeeded by his only surviving son, Louis I the Pious, but long before Louis died his sons were making war over their respective shares of the Frankish inheritance. The principle of unity of the empire, symbolized by the title and supremacy of the emperor and supported by the church, was in conflict with the traditional Frankish principle of equal division of the inheritance of the father among the surviving sons. Louis's eldest son was crowned emperor, but his brothers refused to be deprived of their equal share of territory, and they resisted any effort by their older brother to intervene in their kingdoms. On this basis the Wars of the Three Brothers were brought to an end by the Treaty of Verdun in 843. Three kingdoms, roughly equal in size and resources, were recognized and their boundaries defined. Louis the German, as king of the eastern Franks, received the area generally east of the Rhine and north of the Alps; Charles II the Bald, as king of the western Franks, received the western two-thirds of Gaul; and Lothair I, the eldest son, was recognized as emperor and assigned the narrow strip of territory in between, stretching all the way from the Low Countries through Burgundy and Provence and including the Carolingian kingdom of Italy. In 855 this middle kingdom was partitioned equally among the three sons of Lothair on the latter's death. His eldest son, Louis II, received the title of emperor and ruled Italy; Charles ruled in the kingdom of Provence; and Lothair II inherited the northern third, called Lotharingia (Lorraine), or "Lothair's kingdom." On the death of Lothair II in 870 his uncles, Louis the German and Charles II the Bald, partitioned Lorraine by the Treaty of Mersen. This partition, however, pleased neither side, and the successors of Louis and of Charles waged many wars in their efforts to reunite Lorraine to one or the other kingdom. The Carolingian kingdom of Provence was partitioned between two non-Carolingian claimants: Boso assumed the title of king in the southern part, while Rudolf was recognized as king of Trans-Jurane Burgundy in the north. The two kingdoms were later united as the kingdom of Burgundy under a descendant of Rudolf.

The later Carolingians were incapable of ruling over large territories. The medieval habit of dubbing rulers with brutally accurate nicknames suggests the level to which the line had fallen: whereas Charlemagne was "the Great," his descendants of the fourth and fifth generations were "the Simple," "the Stammerer," "the Fat," or "the Child." One cause of their downfall was the necessity of granting crown lands and privileges to nobles in return for their support in the endless wars of the period. The local nobility of a particular region were more obedient to their count or duke than to the ineffective ruler whose Carolingian descent gave him the title of king or emperor. (R.S.Ho.)

New barbarian invasions. *The Vikings.* Meanwhile, new barbarian invasions were ravaging western Europe, beginning in the second quarter of the 9th century. When the Viking Age began, Scandinavia was by no means an isolated, unknown corner of Europe. Trading contacts with other people had existed for centuries, and colonization abroad had been started by the Swedes in one of the directions the Vikings would later take. The cause of the explosion of the Vikings into the world is a matter of much controversy. Obviously, the urge for adventure and

Cause of Viking expansion



Viking expansion, 7th to 10th century.
From *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

desire for loot motivated some of the Vikings, but deeper explanations vary. Some historians posit overpopulation in Scandinavia, while others point to a deteriorating climate that made agriculture less profitable and more difficult. On the other hand, new land was broken at home throughout the age, and a limit to new cultivation was apparently not reached. The deterioration of the Mediterranean trade, caused by conflicts between Muslims and Christians, undoubtedly played a role, with the northern route from east to west becoming an important alternative. Other factors were the lack of organized resistance to the Norse raiders and the superior shipbuilding techniques of the Scandinavians, which made possible the Viking ship, shallow, pointed at both ends, easily manoeuvrable in rivers and bays, and powered by sail and oars.

The Viking expeditions went in two general directions—west and east—with the Danes and Norwegians more prevalent in the former and the Swedes in the latter. The Scandinavian states were loose confederations at the beginning of the Viking Age, and the identity of the Vikings as Danes, Norwegians, and Swedes is only approximate. The Viking Age is generally accepted as having begun in 793 with the raid on the English cloister, at Lindisfarne (Holy Island, off northern Britain). Monasteries—isolated, wealthy, and inadequately defensed—provided an attractive target for these early Vikings, who were interested in easily won booty and travelled singly or in small groups of ships. After several decades, however, the raiders began to assemble large fleets and to winter abroad. They established camps at strategic places in the British Isles and, later, in France, from which they could make raids into the interior areas; settlements by the Norsemen followed soon after. By 842 half of Ireland was subject to the Vikings, and 11 years later, Olaf the White was the overking of the land where the kingdoms of Dublin and Waterford were established. The Vikings first wintered in England in 851, and in 865 their armies began the overthrow of the Anglo-Saxon kingdoms of Northumbria, Mercia, and East Anglia. In 878 Alfred the Great ceded the territory north and east of a line from London to the northern edge of Wales (about three-fourths of England) to the invaders. This area of Viking settlement was given the name Danelaw; the Danes were predominant in the southern and northeastern parts of the territory, and the Norwegians in the northwest. For more than a century, the Vikings apparently devoted themselves to peaceful settlement and trading in the areas they had won; but around 1000, a new series of raids began, led by the Danish king Sweyn Forkbeard (Svend Tveskæg), and the English were forced to pay tribute (Danegeld) in the hope of avoiding attacks and plundering. In 1013 the attacks reached a peak; all of England submitted, and Denmark and England were joined in one kingdom. But the distance between the two countries and the lack of unifying interests worked to prevent a real unification; when Sweyn died in 1014, it was only through the efforts of his son Canute (Knud) that England remained in the Danish sphere. Canute succeeded his older brother on the Danish throne in 1018, reuniting the kingdoms, and later efforts led to the inclusion of Norway and the Swedish territory of Västergötland in the Danish Empire. Canute devoted most of his attention to England but also sent English missionaries to Christianize Denmark. With the death of Canute's son and successor, Hardecanute (Hardeknud; ruled 1035–42), the Danish empire dissolved. An attempt by the Norwegian king Harald Hardraade (Norwegian Hårdråde; 1047–66)

The Danish Empire of Canute

to reconquer England in 1066 failed but facilitated the conquest of the island by William I the Conqueror, whose Norman duchy had been formed out of another wave of Nordic expansion.

While Vikings from Norway and Denmark were raiding and, later, settling the British Isles, where they soon became assimilated with the native population, the Norwegians were also travelling to the uninhabited or sparsely inhabited islands west of the Scandinavian peninsula: the Shetland, Hebrides, Orkney, and Faeroe islands; Iceland; Greenland; and, finally, North America. The Shetlands were visited quite early: with a favourable wind, the Vikings could reach them from the Norwegian coast in one day. With the Shetlands as a base, settlers moved on to the Hebrides, Orkneys, and Faeroes and to Iceland, where the few residents, Irish monks, were driven out beginning in 874. Expeditions proceeded to Greenland in *c.* 986, and around the year 1000, a few Vikings reached Vinland (Wineland) on the North American coast; no settlements were made there, however.

Of these settlements, Iceland is the best known because of medieval Icelandic documents and histories. A West Norwegian chieftain, Ingólfur Arnarson, moved to Iceland in 874 and was followed by settlers, mainly chieftains and their retinues. By 1100 the population was between 70,000 and 80,000. Early local government by the chieftains led to the establishment of the Icelandic commonwealth in 930, with a national parliament.

While some of the Vikings concentrated on the islands and England, others raided the Continent. By the end of the 9th century, Frisia was controlled by the invaders, mainly Danes, and there were attacks on Paris in 845 and 885–887. The lower Loire was also ravaged during the 9th century. Around 900 the Vikings attacked the northern coast of France, and in 911 the French king ceded land on both sides of the mouth of the Seine to the chieftain Rollo and his followers, who in turn promised to prevent attacks by other Norsemen. The territory ruled by the Danes received the name Normandy, and settlers moved in along with the warriors. The attacks, beginning in the 9th century, in Germany, northern Spain, the North African coast, and Italy were more sporadic, and their permanent effects were negligible.

Russian settlements

The primary attraction of the eastern route, followed primarily by Swedish Vikings, was the riches of the East, brought to Byzantium and Baghdad along the major East–West trade routes; and because these riches were found in well-established, well-defended states, the primary activity of the Vikings here was trade. Early in the 9th century the Swedes had begun to penetrate into Russia, via the Dnepr River to the Black Sea and Constantinople and via the Volga River to the Caspian Sea and Baghdad; and in the later part of the 9th century, kingdoms (most prominently Kiev and Novgorod) were established by the Swedes in Russia. The native Slavs and Finns in the area were easily conquered; the Swedish settlers were quickly assimilated. The export of Russian and Swedish products (furs, slaves, and weapons) and the import of metals and spices from the Orient brought great wealth to Sweden, and the town of Birka, on an island in Lake Mälaren, became an important trading centre during the 9th and 10th centuries; the town disappeared during the 11th century, when the Viking Age was waning. In an attempt to monopolize the northern trade route, the Swedes captured Hedeby (in southern Jutland) in *c.* 900, which remained in Swedish hands for about 30 years.

Decline of the northern route

By the mid-11th century, the Viking Age was over; the last eruption was Harald Hardraade's unsuccessful attempt to win back England in 1066. In the west, the weak states of the 9th century had been replaced by powers no longer defenseless; in the east, the entrance of the Turks produced unease, and the goods of the Orient no longer flowed into Russia as they had a century earlier. Around 1050 the political connections between Sweden and Russia ceased. At the end of the century, the Mediterranean was reopened to the Orient trade by the First Crusade, and the northern route was no longer necessary or attractive. At the same time, Christianity had been established in the Scandinavian countries, and its teachings were inimical to the bold Viking life-style and to the trade of slaves, a prime export. The rise of primitive national states in Scandinavia during the Viking Age also provided an outlet for talents formerly used in trading, settling abroad, or raiding. (H.En.)

Magyars. Another wave of invasions came not from the north but from the east toward the end of the 9th century. The invaders were the Magyars (Hungarians), who made their first authenticated raid into Moravia in 894. The early history of this people is a matter of dispute; but it is certain that they were in southeast Russia in the 9th century, if not before, and probably between the Don and the Kuban rivers; and they appear at one time to have been vassals of the Khazars. Driven westward by the Petchenegs, they arrived at the mouth of the Danube in 889; expelled by the Petchenegs and Bulgars, they entered Pannonia for final settlement in 895 or 896, under their leader, Árpád. They easily subdued the scattered population of the Central Plain, crushed the empire of Great Moravia in 906, and defeated the German forces gathered to meet them in 907. They were then firmly established in Hungary, although Transylvania was probably not truly conquered until at least a century later.

For the following 70 years little is known of the internal history of the Magyars. Árpád died in 907, and his immediate successors, Zsolt (907–947) and Taksony (947–972), are little more than chronological landmarks. During this period the Magyar horsemen ravaged Thuringia, Swabia, and Bavaria, and defeated the Germans on the Lechfeld in 924, whereupon the German king, Henry I, bought them off for nine years while reorganizing his army. In 933 the war was resumed, and Henry defeated the Magyars at Gotha and at Ried (933). The only effect of these reverses was to divert them elsewhere. In 934 and 942 they raided the Byzantine Empire and were bought off under the very walls of Constantinople. In 943 Taksony led them into Italy, and in 955 they ravaged Burgundy. The same year the emperor Otto I overwhelmed them at the famous Battle of the Lechfeld (August 10, 955). This catastrophe convinced the leading Magyars of the necessity of accommodating themselves as far as possible to the empire. They thus retreated into Hungary, accepted western Christianity, and were at last contained. (Ed.)

Foundation of the medieval states. Political recovery from the disintegration that characterized western Europe after the mid-9th century came slowly and took different forms. In England, as already noted, Alfred the Great organized a successful defense against further Viking invasions and won the allegiance of all Anglo-Saxons. His successors slowly reconquered the Danelaw (the areas under Viking rule) and unified all of England under the West Saxon (Wessex) royal house. The strength of the English monarchy was enhanced by the prestige of military conquest; it effectively controlled the local government of the shires under royally appointed sheriffs; and it enjoyed the loyal cooperation of the nobility. In France the monarchy after Charles II the Bald (died 877) declined in prestige as it proved incapable of defending the kingdom or maintaining order. In the 10th and early 11th centuries political stability was rebuilt on the local level by counts and dukes who ignored the king. The crown itself was contested between the later Carolingians and the rising House of Neustria, whose representative, Hugh Capet, finally, late in the 10th century, brought the crown permanently to his family and thus founded the Capetian dynasty. In contrast with the West Saxon kings of England, the Capetian kings in France had almost no influence over most of their kingdom beyond the small royal domain centred on Paris. They were so weak that they aroused no jealousy or resistance from their nominal vassals, the counts and dukes who willingly gave their allegiance.

The most impressive recovery of political stability occurred in Germany, where the last Carolingian, Louis the Child, died in 911 and the greater lords chose one of the powerful dukes, Conrad of Franconia, to be king, as Conrad I. When he died, the nobles again elected their new king, Henry I, duke of Saxony, who repulsed the Magyars and laid the foundations of the German monarchy. His son, Otto I the Great, succeeded by what was in fact a hereditary claim though the formalities of an election were

Europe and the Byzantine Empire *c.* 1000.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

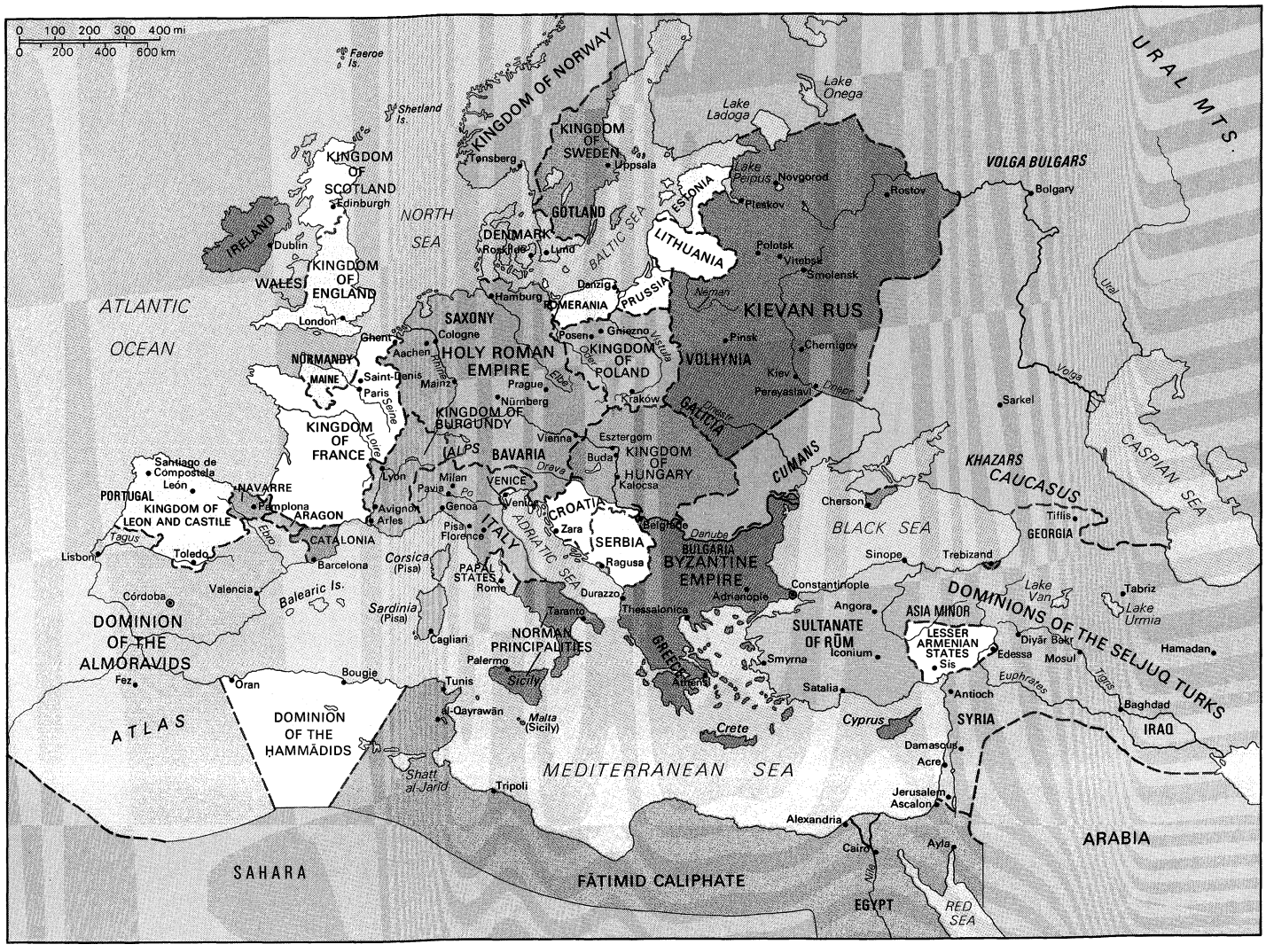
observed. In alliance with the church, Otto I stamped out rebellions and enforced his authority over the duchies and principalities that constituted Germany. In 955 he won a great victory over the Magyars, which permanently ended their raids, and in 962 he made good his claim to be king of Italy and was crowned emperor by the pope. The imperial title had been vacant since 924, and this coronation marks the real beginning of the medieval empire that in a later age was called the Holy Roman Empire.

Defeat of the Magyars

12th-century revival. *The empire and the papacy.* The medieval empire reached its height under Henry III (1039–56), the greatest emperor of the Salian dynasty. Bohemia, Poland, and Hungary acknowledged his suzerainty, and he made his direct rule effective in the kingdom of Burgundy, which his father had annexed to the empire. In Germany he based his power on extensive crown lands, on an imperial civil service recruited from the low-born instead of nobles, and on the church, which he firmly controlled while he promoted ecclesiastical reform. Within the church, reform was sponsored primarily by the monks of the monastery of Cluny and its daughter houses. The Cluniacs welcomed the cooperation of secular rulers, even when such help also entailed secular influence over the election of bishops and abbots. The papacy was untouched by the Cluniac reform and had fallen on evil days, when the throne of St. Peter was the prize of local Roman noble factions. In 1046 no fewer than three men claimed to be pope, against each of whom various charges ranging from simony to heresy had been brought. Henry III intervened: at the Synod of Sutri (1046) he deposed all three and later had his own nominee elected. Under Henry's patronage succeeding popes supported the Cluniacs both in Italy and northern Europe, thus making papal influence effective north of the Alps for the first time since the days of Pope Nicholas I (858–867).

A new phase of ecclesiastical reform was inaugurated by Hildebrand, who became Pope Gregory VII. The Gregorian Reform Party accepted the goals of the Cluniacs but added two more: independence from secular interference in the church and papal leadership of reform. Both issues crystallized in the Investiture Controversy carried on by Gregory and his successors against the new emperor, Henry IV (1056–1106), and his successors. A compromise settlement was finally worked out in the Concordat of Worms (1122), which limited the influence of lay rulers over episcopal elections and prohibited lay investiture of ecclesiastical office.

The feudal monarchies. In England the Norman Conquest of 1066 swept away the upper ranks of the Anglo-Saxon nobility and introduced a new feudal nobility from Normandy and northern France. Uniting the strength of the old English monarchy—the king's peace, an advanced system of local government, and the power to tax—with the powers that he held as feudal lord of the whole realm, William I the Conqueror was the most powerful ruler in western Europe in the most thoroughly feudalized kingdom. Although Normandy and England were divided between his two elder sons, his third son, Henry I, reunited the duchy and kingdom. Because constant warfare on the continent, to protect his duchy, necessitated his frequent absence from England, Henry developed an advanced government, including a specialized department devoted to finances (the Exchequer), which could operate under a viceroy, the chief justiciar.



Europe and the Mediterranean about 1097.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

In France the earliest sign of an increase of royal authority appeared under Louis VI (1108–37), the first Capetian to exercise full control over the royal domain and to win the respect of his great vassals. Though unsuccessful in his attempt to wrest Normandy from Henry I of England, he achieved a greater triumph by marrying his heir to the heiress of the duchy of Aquitaine. When Louis VII succeeded to the throne he was the first Capetian king to rule directly a greater territory than any of his vassals.

The Crusades. In the two centuries after Charlemagne, when western Europe was disintegrating because of civil strife and barbarian invasions, civilization in the Muslim and Byzantine lands flourished. But at just the time when feudal Europe began to regain social and political stability, the once-brilliant caliphate of Córdoba, the ʿAbbāsid caliphate in the east, and the Byzantine Empire under the Macedonian dynasty were all declining.

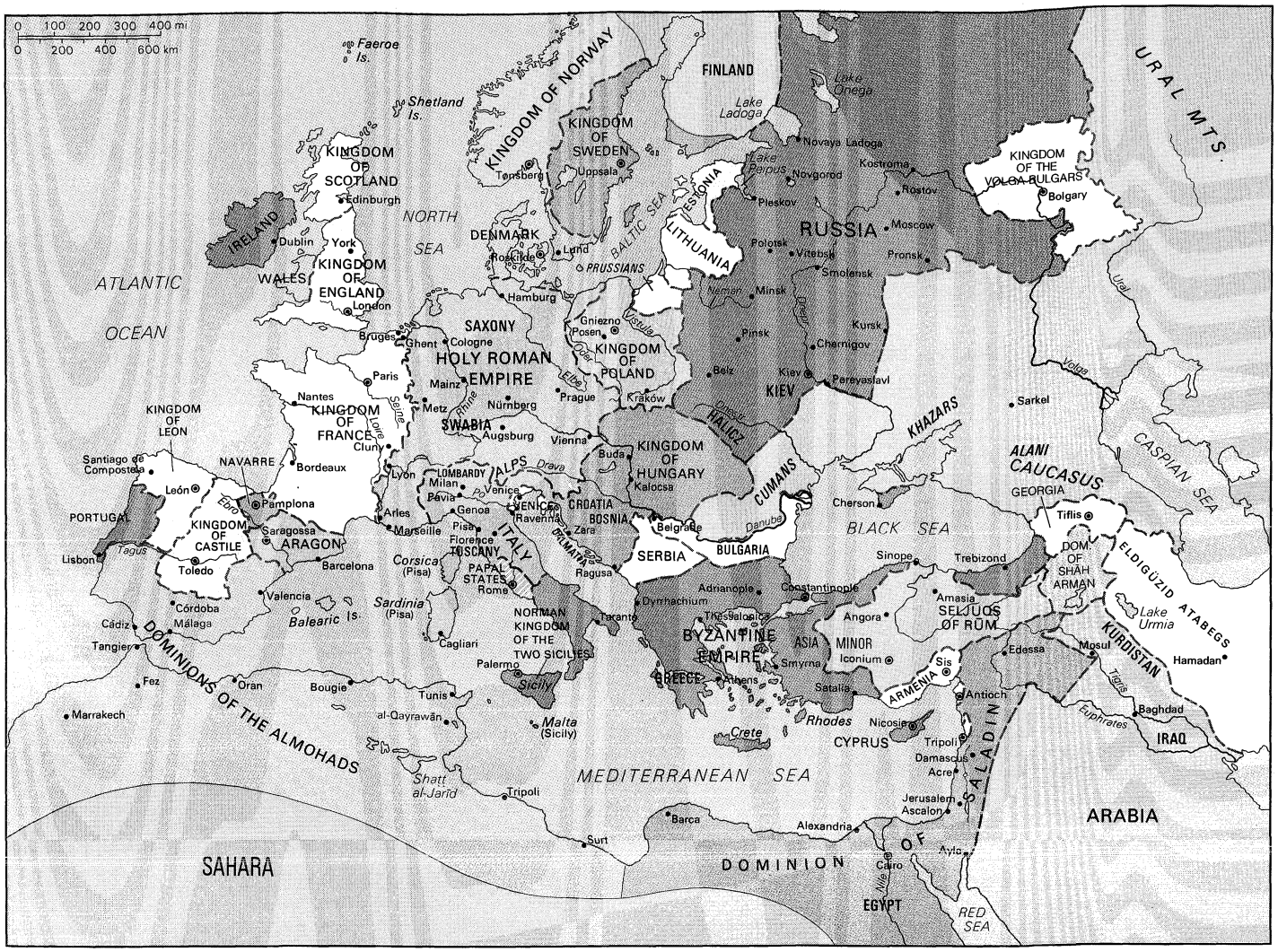
Under such circumstances the petty Christian rulers in northern Spain began to push back the Moors in a centuries-long struggle, the Reconquista, from which the kingdoms of Castile and Aragon emerged as the dominant powers. In southern Italy—an unstable patchwork of principalities under Muslim, Byzantine, and Lombard princes—adventurous and unruly knights from the duchy of Normandy sought their fortune and under the lead of Robert Guiscard won control of both the southern mainland and the island of Sicily, which became the Kingdom of Sicily in the 12th century. In the east Muslim weakness made possible the conquests of the Seljuqs, who first replaced the ʿAbbāsids and then crushed the Byzantine forces in the great Battle of Manzikert (1071). After this victory the Seljuqs swept over Asia Minor and reduced the Byzantine Empire to its European provinces.

The rise of the Seljuqs

To regain his lost provinces in Asia Minor the Byzantine emperor appealed to the West for help, first to secular rulers without success and then to the pope—promising the reunion of the Greek and Latin churches under papal supremacy as the *quid pro quo* by which the pope, Urban II, undertook to raise mercenary troops to reinforce the Byzantine army. At the Council of Clermont (1095) Urban II preached the Crusade. The Emperor and the kings held back, but the greater and lesser nobility of France and Germany enthusiastically responded to the appeal, some feeling a religious dedication to the liberation of the Holy Land from Muslim domination, others welcoming the opportunity to conquer lands for themselves. This First Crusade was successful: Jerusalem was delivered from the infidel, and the crusaders carved out feudal principalities from Muslim Syria and Palestine.

But the Byzantine emperor's hopes for reconquest of his lost provinces were dashed and the compact between the Emperor and the Pope was not fulfilled.

Holy Roman Empire and the apogee of the medieval papacy. After the death of Henry V in 1125, Germany suffered from a disputed succession and civil strife that weakened the monarchy. Peace was reestablished when Frederick I Barbarossa resolutely set out to restore his regalian rights over the lay nobility and the church in Germany and his imperial authority in Italy. Successful in Germany, his efforts south of the Alps led to stout resistance from the Lombard League of northern communes and a revival of the struggle against the papacy.



Europe and the Mediterranean about 1190.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

After their victory at the Battle of Legnano (1176) the Lombard communes secured almost complete rights of self-government in return for recognition of the emperor's suzerainty. With the Pope, Barbarossa came to a compromise that settled none of their conflicts; on balance, it was a defeat of the Emperor's effort to control the papacy and church in Italy as he did in Germany. Papal independence of secular interference was further assured by a decree of the third Lateran Council (1179) that a two-thirds majority of the College of Cardinals sufficed in the election of a new pope, without confirmation by the emperor. Barbarossa's only success in Italy was the marriage of his son to the heiress of the Kingdom of Sicily.

Under Pope Innocent III (1198–1216) the papacy reached the height of its prestige, leadership, and power, and the ideal of the unity of Christendom came nearest to realization. To the papal curia appeal lay from all the ecclesiastical courts of Europe, and the courts of archdeacons, bishops, and archbishops, steadily enlarging their jurisdiction since the days of Gregory VII, entertained accusations or disputes involving not only all the clergy but all the laity who were *miserabiles personae* (widows, orphans, paupers, and others incapable of protecting themselves). The ecclesiastical courts also claimed jurisdiction over all disputes concerning the discipline or administration of the church, property claimed by clergy or ecclesiastical corporate bodies, and questions touching marriages, wills, vows and oaths, the sacraments, and heresy. Over all these matters the pope was supreme judge (*judex ordinarius*) and even claimed the right to interfere in temporal affairs where sin might be committed and the salvation of Christian souls endangered—as, for example, when war threatened between the kings of England and France, Innocent intervened as peacemaker in a wholly feudal quarrel. The fourth Lateran Council, summoned by Innocent III in 1215, legislated on matters of church reform; condemned the use of superstitious ordeals in all courts, both ecclesiastical and lay; proclaimed a Crusade; deposed the emperor Otto IV in favour of Frederick II; condemned the heresy of the Albigensians and excommunicated their lay protector, the Count of Toulouse; required all Christians to partake of the sacrament of the Eucharist at least once each year; and pronounced transubstantiation to be a dogma.

Heresy flourished mostly in towns, where ideas as well as commodities from far-off places were traded and where the growth of the population far outstripped the increase of churches and clergy. To minister to these needs and to meet the threat of heresy two new monastic orders were established: the Franciscans and the Dominicans, who, instead of retiring from the world, preached in the towns. Where heresy was so strongly entrenched that it had to be repressed, the special papal ecclesiastical court, the Inquisition, was employed. Dominicans staffed the court; lay rulers were obliged under pain of excommunication to carry out its sentences.

The Franciscan and Dominican orders

Frederick II (1215–50), who inherited the Kingdom of Sicily from his mother and the Holy Roman Empire from his father, renewed the struggle with the papacy. From the conflict both sides emerged exhausted. In Germany imperial authority almost lapsed; in Italy the Lombard towns staved off Frederick's efforts to unite all Italy under his autocratic rule, and the resources of Sicily were drained

off by war. The papacy lost prestige by proclaiming a Crusade against a Christian monarch and employing spiritual weapons (excommunication or interdict) for worldly ends. The 13th-century popes were able administrators and dedicated reformers, but they failed to preserve the spiritual leadership. After Frederick II's son Conrad IV there was a period of disputed succession and vacancy in the empire, the Great Interregnum (1254–73) that witnessed the collapse of central authority. Yet the greater prelates and princes of the empire maintained order within their archbishoprics and bishoprics or their duchies, counties, and other principalities.

Rise of the Western monarchies. In England, France, and Spain the medieval monarchs founded national states that were to survive—in contrast with the more universal and supreme claims of both papacy and empire, neither of which could preserve the spiritual and temporal authority exercised by earlier popes and emperors. In England the state-builders who accomplished most were Henry II and Edward I. In France, Philip II Augustus (1180–1223) by his conquest of Normandy from John of England not only initiated the growth of the royal domain, which finally included all the kingdom under the king's direct rule, but also precipitated the quarrels between John and his barons that led to the issue of Magna Carta (1215). If Philip II won the respect, Louis IX (1226–70) won the love of Frenchmen. St. Louis made the monarchy the focus of national aspiration and pride in France, and his successors made good use of the loyalties he had awakened.

In the 14th and 15th centuries England and France fought the Hundred Years' War, a conflict in which the English lost all their French possessions (except the port of Calais), and France was ravaged by the pillaging and dislocations of war. Yet each monarchy emerged stronger at the end. Emergency conditions in France allowed the kings to assume the power to tax all subjects without consent and to maintain a standing army of professional soldiers. The continual need for money to support the military effort had the opposite effect in England, where Parliament assumed more power because the kings had to gain its consent to taxation. The aftereffects of the war disrupted England and led to the dynastic Wars of the Roses, the only solution of which was acceptance of a strong monarchy. In Spain, the Reconquista was brought to a successful climax under the lead of the kings of Castile and Aragon. After 1270 only the little kingdom of Granada remained unconquered. The marriage of Isabella of Castile and Ferdinand of Aragon in 1469 was followed by the union of Castile and Aragon and the expulsion of the Moors from Granada (1492). Ferdinand and Isabella employed the Inquisition to drive out Moors and Jews and to eliminate political opponents by convicting them of heresy.

Eastern frontiers of medieval Europe. At the times of the great migrations (*Völkerwanderung*) the Slavs—referred to by Tacitus, Pliny, and Ptolemy under the name of Veneti or Venedi—lived between the middle Vistula River to the west, the Carpathian Mountains to the south and the Dnepr River to the east. In the north their neighbours were the Balts, ancestors of the Old Prussians, Lithuanians, and Letts, or Latvians. After the departure of the Goths and other Germanic tribes (who moved southward, entering the territories of the Roman Empire), the Slavs advanced westward, crossing the Elbe-Saale line in the 5th century, and southward, penetrating into the Danubian Basin and the Balkan Peninsula. Three centuries later, when Charlemagne conquered and evangelized the lands of the heathen Saxons, the western Slavs found themselves in danger. On the borders of a great empire with a superior culture, they led a pastoral life with a rudimentary social and political organization. A German *Drang nach Osten* (drive to the east) began. The idea of a universal Christian empire inspired it; demographic and economic conditions gave it force. The density of population was greater west of the Rhine than east of the Elbe, and by the end of the 8th century there was little unsettled agricultural land within the Frankish state. The Slavonic tribes living between the Elbe and Oder rivers—the Polabs (the Obodrites, or Bodrycy, and the Wilcy, or Lutycy) as well as the Lusatian Sorbs—were partly conquered during the 9th century, their conquerors being mainly the Saxons, who had become the outpost of German eastward expansion.

Poland. To protect them against the fate that had befallen the Polabs and to enable them to defend themselves against the German advance, the Piast dynasty during the 9th century grouped the Polish tribes—Polans, Mazovians, Vislans, and Silesians—into one state. The fifth prince of this house, Mieszko (or Mieczysław) I, received Christianity direct from Rome in 966, thus depriving the German emperors of a pretext for a missionary drive into Polish lands east of the Oder. His son, Bolesław I the Brave, in agreement with the emperor Otto III, secured the independence of the Polish church (1000) with an archbishop at Gniezno; in 1024 Bolesław was crowned king of Poland. But this did not stop the German pressure. More than a century later, in 1107, Archbishop Adelgot of Magdeburg appealed to his compatriots to carry the Christian faith to the Poles in these words:

The Polish church

> They are the worst of pagans, but their land so abounds in the best of meat, honey, corn, and all products of the earth that no other land may be compared with it. Wherefore, O Saxons, Franks, Lotharingians, and Flemings, you can there both save your souls and gain the best of land in which to live.

For two centuries the Oder River was the western frontier of Poland, but in 1135 King Bolesław III the Wrymouth had to acknowledge the emperor's suzerainty over western Pomerania. Bolesław died three years later. He divided his kingdom among his four sons, Władysław, the eldest, receiving Silesia. His position as suzerain was enhanced by his tenure of Kraków, the capital, and of Pomerania (Pomorze). By the beginning of the 13th century the Old Prussians and Lithuanians were the only heathen neighbours of Poland. The former resisted conversion by Polish missionaries and often raided Poland instead. Prince Conrad of Mazovia, whose lands suffered most, asked the Teutonic Order in 1225 to help in evangelizing Prussia. The offer was accepted, and a few decades later Poland had a powerful German state on its northern border. In 1308 the Teutonic Order conquered Pomerania and Danzig. Władysław I the Short reunited the kingdom in 1320 and was crowned king in Kraków but was unable to dislodge the order from eastern Pomerania. As the western part of this province was transferred in 1225 from the Polish archbishopric of Gniezno to that of Magdeburg, a German corridor separated Poland from the Baltic. This sea became a German lake for two main reasons: first, because by 1227 the present-day Latvia and Estonia had been conquered by the Order of the Knights of the Sword (which in 1237 joined the Teutonic Order); second, because the Hanseatic League, founded in the mid-13th century by the German maritime towns to control trade between eastern and western Europe, had become a power able to wage war against Denmark and dictate to King Valdemar IV the Treaty of Stralsund (1370). At that time more than 50 towns—from Brugge to Reval (Tallinn)—belonged to the league, and the Hanseatic settlements in the east helped the process of Germanization.

The Mongol–Tatar invaders who devastated southern Poland in 1241 were stopped and defeated at Legnica by the army of Prince Henry II the Pious of Silesia (of the House of Piast), who perished in the battle. To rehabilitate the country, his successors encouraged the immigration of German peasants and artisans, which helped the country's economic and cultural advance but also contributed to its Germanization. The growing German menace led to a Polish–Lithuanian union under the Jagiellon dynasty. Their combined forces defeated the Teutonic Order at Grunwald and secured the return of Pomerania and Danzig to Poland (1466). Poland became a great power in eastern Europe and the *Drang nach Osten* was brought to a halt.

Bohemia. Great Moravia, the first western Slavonic state, was founded at the beginning of the 9th century. It included not only the present Czech and Slovak lands but also Lusatia, the southern part of present-day Poland, and the western part of present-day Hungary. Great Moravia was weakened by German invasions from the west, and, at the beginning of the 10th century, it was destroyed by the Magyars.

In Bohemia proper the Czech Přemyslid dynasty founded in the first half of the 10th century a state the first historically known ruler of which was Wenceslas (Václav), assassinated by his brother in 929. In 1212 Bohemia became part of the Holy Roman Empire, and German colonization followed. The Přemyslid dynasty, extinguished in 1306, was succeeded by the House of Luxembourg. The second ruler of this dynasty—king of Bohemia as Charles I and emperor as Charles IV—resided in Prague, raised its bishops to the rank of archbishops (1344), and founded Prague University (1348). In 1356 he also fixed the attributions and the composition of the imperial college of electors, including the king of Bohemia as one of the four lay members. (The three others were the count palatine of the Rhine, the duke of Saxony, and the margrave of Brandenburg; the bishops of Mainz, Trier, and Cologne were the ecclesiastical electors.)

Before Charles I's reign the Catholic hierarchy of Bohemia and its bourgeoisie were German. A Czech bourgeoisie now appeared on the scene, and that marked the beginning of a struggle against German preponderance in the country's cultural and economic life. The movement for church reform headed by John Huss was to a great extent a national one. The burning of Huss (1415) was in the eyes of the Czech people an act of collusion between the Germans and the pope. A civil war started in Bohemia, and it ended with the election of George of Poděbrady (1457), a Czech nobleman, as the Hussite king of Bohemia.

John Huss

Hungary. The Magyars, after the destruction of Great Moravia, subjugated Slovakia and settled the Danubian Plain. Their raids against Germany ended after the defeat they suffered at Lechfeld (955). Géza, a ruler of the Árpád dynasty, was converted to Christianity in 975, but it was under the reign of his son, Stephen I, that Hungary, a multinational state, entered the European community, as a kingdom outside the Holy Roman Empire. In 1102 Hungary and Croatia were united under St. Stephen's crown. The Mongol invasion of 1241–42 left the country ravaged and depopulated. When the House of Árpád became extinct in 1301 it was succeeded by the Angevins (1308–82)—a period of power and prosperity for Hungary. In 1345, however, the Ottoman Turks had secured a bridgehead in Europe. In 1363 an allied army led by King Louis I of Hungary, was defeated by the Turks on the Maritsa River in Thrace. In 1389 the Serbian Empire founded by the Nemanja dynasty was destroyed by the Turks at Kosovo. Hungary itself was in danger.

After Louis I's death Hungary had as kings a Luxembourg (Sigismund) and a Habsburg (Albert), who, being elected Roman emperors in succession, were concerned more with the interests of the empire than with those of Hungary. In 1440 Władysław III Jagiełło, king of Poland, was crowned as King Ulászló I of Hungary, but in 1444 he fell at Varna, leading a Hungarian-Polish army against the Turks. János Hunyadi, the Hungarian commander who survived the Varna disaster, became regent of Hungary for Albert Habsburg's posthumous son, King Ladislas V. When Ladislas died in 1457 at age 18, Matthias Corvinus became the only Hungarian national king after the Árpáds.

Scandinavian kingdoms. The pagan Norsemen discovered toward the 8th century that plundering rich foreign lands was more profitable than extracting a living from the sea or the unfertile soil. While the Swedes turned their attention to the east of the Baltic and gave Russia its first dynasty, the Danes and Norwegians moved westward to Scotland, England, and Ireland and also southward to France, Spain, and Sicily. Under Canute the Great (*c.* 995–1035) Britain was part of a Scandinavian empire comprising Denmark, Norway, and part of Sweden. Canute became a Christian and supported the evangelization of Scandinavia. After his death his empire fell to pieces. Denmark, Norway, and Sweden became independent kingdoms warring among themselves. Between the 12th and 14th centuries Denmark was the leading Scandinavian power. For a time it annexed Estonia, the then Slavonic island of Rügen (Rugia), and part of Pomerania. Sweden undertook in the 12th century the evangelization and conquest of Finland and in 1323 concluded a peace treaty with the Russian duchy of Novgorod that fixed the first eastern frontier of Finland.

In 1397 a union between Denmark, Sweden, and Norway was concluded at Kalmar. It was ruled by Margaret I, daughter of a king of Denmark and widow of a king of Norway. The Swedish nobles banished their unpopular king, Albert of Mecklenburg, and elected Margaret as "sovereign lady and ruler." She died in 1412 and was succeeded by Erik of Pomerania. His plan to create a great Baltic empire did not materialize, and the Hanseatic League continued to control the Baltic trade. The Kalmar Union continued formally until 1523, although in Sweden there was strong opposition to it.

Russian duchies under Mongol domination. The invasion of Russia in 862 by Rurik (Hrörekr) and his Norsemen marks the beginning of Russian recorded history. The Norsemen followed the rivers and lakes from the Baltic to the Black Sea and established their capital in Kiev. Vladimir Svyatoslavich, grand duke of Kiev (978–1015), probably a direct descendant of Rurik in the third of fourth generation, accepted the Christian faith from Byzantium in 988. There were many Russian duchies and endless family wars waged with the purpose of deciding which among the large number of Rurikovichi was at the time the senior and thus entitled to rule in Kiev, the richest duchy. The Mongol-Tatar invasion of 1237–40 destroyed the Kievan power and subjected all Russian lands (except Novgorod) to the domination of the Golden Horde. In spite of Dmitry Donskoy's victory at Kulikovo (1380) it lasted until 1480.

The church in the later Middle Ages. The most extreme claims to papal power and ecclesiastical exemption from royal jurisdiction were made by Boniface VIII (1294–1303), who forbade taxation of the clergy without consent of the pope. This precipitated a contest between Boniface and the kings of England and France, each of whom claimed that he had to tax his clergy to defend his realm against attack from the other king. In a brief struggle, replete with vituperative propaganda against Boniface, the monarchs were triumphant, and the pope was reduced to claims he had no means of enforcing. His successors, French popes, resided at Avignon from 1309 until 1377, when Pope Gregory XI returned to Rome, where he died. The cardinals then elected an Italian as the new pope, Urban VI, but then—claiming that election to be invalid because it was made under duress from the Roman mob—elected a new French pope, Clement VII, who promptly returned to Avignon. Thus began the Great Schism (1378–1417) that divided the allegiance of Europe about equally, while each pope excommunicated the other and the other's supporters. This scandalous situation led to demands not only for reunion but for reform of the papacy as well. The great councils held at Pisa (1409), Constance (1414–18), and Basel (1431–37) restored unity by electing Martin V (1417–31) and deposing all other claimants, of whom there were then three. The conciliar movement failed to achieve significant reforms or to limit papal authority. The price paid for papal victory was high: to woo away the support of secular rulers the popes made concessions that amounted to sharing papal authority with them. But even more costly was the loss of prestige, which the Renaissance popes did nothing to make good.

The Avignon papacy

Fall of Constantinople. By the end of the 13th century the last of the Crusaders' principalities had fallen to Muslim reconquest, despite the efforts of such leaders as Frederick Barbarossa, Richard the Lion-Heart, Frederick II, and St. Louis, each of whom led a Crusade without lasting result. In 1291 Acre, the last city of the kingdom of Jerusalem, fell, and only the island of Cyprus remained in Christian hands. In the 14th century the Ottoman Turks swept westward, attacking Muslim, Byzantine, and independent territories in both Asia and Europe. They crushed the greatest Balkan power, Serbia, and overran the rest of the peninsula. They were then free to complete the conquest of the rest of Asia Minor and to lay siege to Constantinople. The great fortress city fell in 1453. This event hardly changed the situation in the Levant—Italian traders continued to enter the ports and markets, though Venetian trade suffered.

Italy and Germany in later Middle Ages. Both north and south of the Alps political disintegration of the Holy Roman Empire was accompanied by the growth of smaller states. In Italy these were the city-states, dominated by the wealthy merchants. Venice, Milan, Florence, and Rome were the most important, while the Kingdom of Naples was still a power. Though many of the city-states had republican forms of government, almost all were in fact governed either by a small oligarchy or by a single despot. The popes, in their policies and their plots and intrigues, were hardly distinguishable from other despots. In Germany the situation is summed up in the term particularism—*i.e.,* the division of the country into autonomous principalities and self-governing free cities. The emperor was elected by seven electoral princes, and after 1356 papal claims to adjudicate a disputed election were explicitly denied. After 1438 the Habsburg dynasty once more came to the imperial throne, this time to retain the title permanently. Socially and culturally Italy and Germany present a striking contrast at the end of the Middle Ages. In northern Europe society was still dominated by the rural nobility, in Italy by the urban bourgeoisie. In arts and letters, as in education, medieval styles, interests, and methods continued in the north, while in Italy the Renaissance brought a revival of the literature, artistic styles, and interests of classical antiquity.

(R.S.Ho./Ed.)

MEDIEVAL SOCIETY

For most of the medieval period in most of Europe, the structure of society was determined by the difficulties of providing an adequate and continuous supply of food and raw materials. The proportion of grain reaped to seed sown was generally low; acute difficulties of transport made agricultural specialization hazardous and prices local and unstable. Hence, particularly between the 8th and 11th centuries, the proportion of the population freed of all agricultural tasks was low, and the materials of luxury or warfare were rare and highly prized.

The aristocracy that could be supported by such a society was both in form and origin a blend of Roman and Germanic elements. Late Roman society was marked by the existence of an aristocracy of wealthy landowners whose estates were worked by slaves either maintained in the household or housed on small plots about the great house. Other dependents, of rather higher status, might be freedmen or formerly independent cultivators, such as the *coloni,* whom war or financial distress had brought under the landowner's protection. Well before the formal end of the empire in the West, such landowners had been accustomed to deal out a usurped justice even over their free clients and had maintained armies of *bucellarii* (bodyguards) in their own defense, though rarely placing a high value on military accomplishments. With difficulty they retained their contacts with the towns and with the traditional urban education of their class.

Decline of the towns

Earlier Roman government had depended upon these towns, which were almost universally in decline by the 5th century. Vandal and, later, Muslim piracy disrupted the vital sea routes to Africa and the East; on land the impotence of local government made communications dangerous; and ever heavier taxation crippled trade. Long-range commerce and the more local urban industries declined, until even wheel-turned pottery became scarce or vanished. The retreat toward economic self-sufficiency and a barter economy was reflected in the near collapse of Western coinage. Over this society, rapidly receding into a pre-Roman localism, there presided, with ever-feebler effect, an emperor or emperors who attained office at the head of an army and according to no settled principle of succession. Once in power the emperor exercised a nearabsolute authority over the army, the courts, and the administration, limited only by prudence and the likelihood of mutiny or assassination. Emperors were hailed as themselves divine before the conversion of Constantine early in the 4th century; thereafter, even a Christian emperor who enforced the precepts of the church readily took on a quasi-priestly character.

The Germanic invaders who settled in or along the frontiers of this empire lived very differently. Not all seem to have had kings, at least permanently, but the majority, whether as allies or invaders, entered the empire as warrior bands led by chiefs who depended for power on their continued capacity to win battles and to retain followers with gifts of present wealth and hope of more. Kings were ring givers, holders of great feasts, lords of men. Yet their authority rested on other grounds too—their descent from the gods and a belief sometimes current that divine favour was assured the people through the royal lineage. Some of the pagan kings (such as the Ynglings of Sweden) acted as priests, and a king's subjects usually followed him to the font for baptism as they followed him to war. Very special measures, including the earliest recorded Frankish consecration by a bishop, were required to legitimize the succession of a king from outside the ruling dynasty.

The politically active element in the people was represented by the warriors—the freemen—who formed the army; when this was gathered it represented the people for war or peace, and it was the army that acclaimed each new king by raising him upon their shields or enthroning him upon some sacred stone. Among these freemen there was a group of special importance and standing, set apart by their birth and their lord's favour—the *comitatus,* or *gasindi,* already known to the Roman historian Tacitus in the 1st century AD. For them, loyalty to their lord was the supreme virtue, celebrated in epic verse from *Beowulf* on. From their ranks were drawn the barbarian officers of post-Roman and Germanic society, though the highest such posts seem to have been monopolized by a restricted number of noble kin groups, distinguished by their descent and deeds rather than by any set title to property.

The rulers. *Kings.* In the societies in which these disparate patterns blended, beginning in the 8th century, the medieval kings were raised to office by a combination of ritual acts that revealed the elements of their power. Until the principle of male primogeniture became dominant in the late 12th century, the king would first be "chosen" in an assembly from among the kin of the last ruler, whose designation would carry great weight in cases of doubt; this choice was often not complete until the new king had travelled through his kingdom in what has been called a continuous election. After the consecration of Pepin the Short as king of the Franks by St. Boniface in 751, it became increasingly common for the king so acclaimed to be consecrated with a liturgy carefully modelled upon the Old Testament precedents of Saul and Solomon and to be invested with such insignia of kingship as crown, sword, helmet, or sceptre. Before or after this ceremony, the leading men of the kingdom came in to declare their allegiance, often performing symbolic acts of domestic service at the coronation feast. In these ceremonies the king secured both rights and duties. As the anointed of the Lord, he had a special claim on the obedience of the church and a measure of physical security; the murder of Canute II of Denmark in 1086 horrified Europe, and the violent deaths of kings continued to be regarded as striking at the whole fabric of the divine and human order. This process, whereby the church sharply distinguished the king from the chieftain, in part explains an important transition of the 9th and 10th centuries. Thenceforward, the multiplication of kings, common among the descendants of Clovis or Charlemagne, ceased; only a formal act of the papacy under rare conditions was thought capable of legitimizing such later kingdoms as Hungary (999), Sicily (1139), or Portugal (1143).

Medieval notions of kingship

The church, which played so large a part in creating such a king, was active in prescribing duties; the coronation oaths and prayers insisted upon his obligation to protect the church, the defenseless, and the poor, to make war upon the heathen in the service of Christ, and to ensure that justice was done. It was chiefly in church councils that the king's duty to the whole people was emphasized against his relations with his warriors.

Even where, as with the Capetian dynasty between 987 and 1316, son succeeded father in unbroken descent, it was conventional to refer to the king as chosen by his people, and this became an active principle when there was no obvious claimant, as at the end of the Saxon or Ho-

henstaufen dynasties in Germany or in such exceptional cases as the founding of the Latin kingdom of Jerusalem in 1099. Among the larger kingdoms of the 13th century, only Germany, beset by frequent changes of dynasty and by the papacy's hostility to a hereditary empire, was still in practice an electoral state. Charles IV's Golden Bull of 1356 formally defined the procedures of election that remained in force in the empire until 1806, although after 1437 the throne was monopolized by the family of Habsburg. In Denmark, Sweden, and Poland the kingship was also formally and often practically elective, though hereditary right became absolute in France, England, and Spain.

Emperors. Theoretically, after 800 the emperors—the heirs of Charlemagne—stood above the kings. The appearance of a second Christian empire caused endless difficulties with Byzantium, and the nature of its pre-eminence was always diffuse and uncertain. It could be considered a recognition of facts—the attribution of a supreme title to the most powerful of the Latin kings—and many saw Charlemagne or Otto I in this light; but none of their successors enjoyed the same authority as did these founders of empire. The more general view, dominant by the end of the 10th century, was that the empire was specifically Roman, but in several possible senses. Because it was Pope Leo III who had taken the lead in recognizing Charlemagne as the first emperor in the West in 300 years, and, because Leo's successors had the undoubted right to consecrate later emperors at Rome, it could be claimed that the empire was an office within the Christian community, or *ecclesia.* The chief duty of such an emperor would then be the protection of the church and the enforcement of ecclesiastical discipline. When the emperor Henry III summoned the synod of Sutri of 1046, which ratified the withdrawal of three rival popes, he fulfilled one interpretation of this role; when Pope Gregory VII declared Henry IV deprived of his kingship in Italy and Germany for offenses against the church in 1076, he was employing another. Rome was, however, the city of the Caesars as well as of St. Peter, and some claimed that the emperors were the heirs of Augustus and the pagan emperors as well as of Charlemagne, with a universal authority that owed nothing to the church. This view was asserted notably by the apologists of Frederick I Barbarossa (reigned 1152–90), supporting their claims upon the revived study of the Roman civil law; modified by the doctrines of Aristotle, it survived in the writings of Dante and among the courtiers of the emperor Henry VII (reigned 1308–13).

Chief duty of the emperor

Without effective control in Rome, all of these views lacked substance, and such control was made impossible because of political stress in Germany under Henry IV and, later, the rise of the Lombard towns (which inflicted a serious defeat on Frederick I at Legnano in 1176) and the papacy's determination to establish an independent state in central Italy under Innocent III and his successors. After the death of the emperor Frederick II (reigned 1215–50), who had united the crowns of Germany, Italy, and Sicily in his own person, prolonged succession disputes on both sides of the Alps ruined the foundations of central authority in Italy and Germany. After the reign of Otto I (German king, 936–973; emperor from 962), the king of Germany had an exclusive title to the empire, though he could only secure it through coronation at Rome. The universal and ecclesiastical pretensions enshrined in this act had become a fiction long before the last imperial coronation in Rome, that of Frederick III in 1452, the bulk of whose reign was spent trying to maintain a foothold even in his own duchy of Austria.

Popes. If the imperial authority became progressively more confined, that of the popes made great advances, partly at the empire's expense. The foundation of papal authority lay, first, in the claim to inherit all the powers conveyed to St. Peter by Christ and, second, in the special position of the bishop of Rome. Until the 8th century, papal pre-eminence was complicated by the existence of jealous patriarchates of the East, not least the new Rome at Constantinople, whose emperor remained the nominal and sometimes the effective overlord of the city of Rome. By 700, however, most of the other patriarchs were subject to Muslim rulers, and Byzantine military weakness had forced such popes as Gregory I to take on most civil responsibility for the city of Rome. Dogmatic disputes in the Byzantine Empire over the veneration of icons (725–843) hastened a process of estrangement between the Greek Eastern and Latin Western churches that was to deepen eventually into near permanent schism. Rome then was left in solitary eminence in the West. Missionaries from Anglo-Saxon England, which had been converted by Roman agents, came to exercise a dominant influence upon the Frankish kings of the early 8th century; under Pepin the Short, Charlemagne, and the latter's son Louis the Pious, strenuous efforts were made to enforce a single liturgy, canon law, and monastic observance, the authenticity of which was guaranteed by its use at Rome.

Until the 11th century this confirming and legitimizing of action begun elsewhere was the predominant if not exclusive role of a papacy largely under the control of a series of such local Roman noble houses as the Crescentii and counts of Tusculum. The papacy favoured rather than led movements of active reform in the monastic and secular church. With the accession of Gregory VII, however, the papacy assumed the leadership of a movement that saw the inertia of the bishops and archbishops as the chief obstacle to ecclesiastical reform and viewed this situation as a consequence of excessive lay influence on their appointment and conduct. Accordingly, Gregory and his successors pursued a policy aimed at reducing the authority of the secular princes over the church (so provoking the long and bitter Investiture Controversy with the emperors Henry IV and Henry V) and at replacing secular authority with an ecclesiastical government based upon the written canon law and directed from Rome. By the pontificate of Innocent III (reigned 1198–1216) substantial successes had been achieved. The Fourth Lateran Council of 1215, attended by more than 400 bishops and 800 abbots from the whole of the Latin obedience and by representatives of all the greater princes, disposed of secular and ecclesiastical business on the widest scale. The difficulties that beset a full realization of the papal vision of a Christian and priestly monarchy were both internal and external. Internally, the process of centralization of authority made for slow and expensive decisions and was thought to demand a political independence in Italy that, in turn, involved long and costly wars with the Hohenstaufen and their successors north and south of the Papal States. Both centralization and the need to levy taxes on the church at large aroused the distrust and resentment of the local hierarchy, which could be exploited by kings (such as Philip IV of France in his contest with Boniface VIII) who were determined to secure their own right to tax their clergy and to submit them to the royal law.

Papal leadership

The rise of national consciousness that accompanied this resistance as well as the exile of the popes at Avignon made the pope's universal role as arbitrator in secular or ecclesiastical affairs harder to sustain. The double election of 1378 created a schism and weakened the position of the papacy. Because neither contestant would give way, the schism could only be ended by an external agency—a general council of the church held at Constance in the presence of the emperor Sigismund between 1414 and 1418. The energies of the abler reformers of the next 30 years or so were largely diverted to disputing the rival claims to authority of a monarchical papacy and a general council. With the failure of the most ambitious of these at Basel, the papacy emerged from the struggle again in Rome and again enjoying the plenitude of power; but the chief beneficiaries of the struggle were the secular princes, who had secured valuable concessions of control over their clergy as the price of their support.

The aristocracy. *The greater aristocracy.* By 1100 a greater aristocracy had evolved over almost all the Latin West, marked by a combination of three elements. First, its members were normally the lords of a number of men bound to them by an oath of fealty and an act of homage, while they were themselves so bound to a king, a prelate, or the pope; the obligations such bonds involved were various, but the performance of military service was the most widespread and characteristic. Second, this aristocracy commonly exercised a large measure of judicial,

Elements of aristocracy

financial, and administrative authority over its own men as lords and often over others as royal officers. Third, this was a landowning aristocracy that rewarded its dependents with grants of lands (fiefs), much as its own wealth had been enhanced by grants from kings, and maintained large households from the proceeds of the estates retained in the lord's own hand—the domain.

To this combination, or some elements of it, the term feudalism is commonly though inconsistently applied; Marxists sometimes further apply it to a particular form of agricultural exploitation, the manorial system. But none of these elements necessarily supposes the others: in England it could be said that all land was received from the king, yet public authority remained unusually concentrated in his hands; in Germany the greatest princes came to hold rights of jurisdiction from the emperor rather than land, and their reciprocal duties were very slight; in southern France and Italy tenure for homage and service was rare but political authority was exceptionally fragmented.

Within the ranks of the greater aristocracy, certain distinctions came to exist. Originally, there seem to have been a limited number of very great families in Germanic society, defined by birth and sometimes uniquely described as free, but this division slowly gave way to others whose titles derived from functions rather than social rank. The term duke was widely applied to the lords of certain areas, which were incorporated in the Frankish empire without wholly losing their identity—*e.g.,* Bavaria, Burgundy, Brittany, Aquitaine, Saxony—and also to the lesser lords of the loosely constructed kingdom of the Lombards; the Scandinavian jarls or Anglo-Saxon earls after the reign of Canute (died 1035) were similar. In origin the word duke (*dux*) meant military commander, and it was on the basis of military need that the majority of 10th-century duchies emerged; for similar reasons, there appeared the margraves of Italy and Germany, lords of border land with a wider military authority. In Germany duchies multiplied in the 12th century, as in the creation of Austria in 1156, when the title implied membership of the highest rank in the social and tenurial scale; in 13th-century France and 14th-century England the term was only revived for great lordships (appanages) created for the cadets of the royal family.

The most widespread of such titles originally denoting public office and only later social rank or landed wealth was the Latin *comes* (German *Graf,* Anglo-Saxon *ealdorman*), or "count." *Comes* originally meant companion or member of the king's household of specially trained warriors; the appointment of members of the Frankish king's inner circle to the earlier administrative districts of *gau* or *pagus* caused the word to be applied to an officer of Carolingian government, then to a hereditary holder of office and rights, and finally to the head of a noble landed family. The terms duke, marquis, and count were themselves no necessary guide to relative wealth or prestige because some duchies were almost empty titles while the powerful lords of Flanders or Champagne were only counts. The institution of the *Reichsfürstenstand,* the class of imperial princes in late 12th-century Germany, was conceived as defining the highest rank of lay and church princes standing immediately about the throne. While the French institution was to become essentially an empty honour under the monarchical government of later kings, the German class was buttressed with important privileges that gave it enduring importance.

Recognition of rank

The origins of this greater aristocracy and its privileges were diverse. Some Roman senatorial families survived or absorbed their invaders; some descendants of independent chieftains of war bands that did not secure lasting kingdoms entered the Frankish, Anglo-Saxon, or Scandinavian nobility; the descendants of some specially favoured royal companions were able to hold onto the gains of their ancestors. Birth was from the outset probably an essential element of both exalted rank and access to royal favour, but important modifications to this took place between the 9th and the 12th centuries; slowly, the emphasis upon patrilineal descent and primogeniture became ever more absolute, and the extended kin group gave way to the dynasty. The passing of public offices into hereditary possession, the growing importance of military considerations, and a concentration of wealth and influence upon a limited number of indivisible castles all contributed to this effect. Whereas the great Carolingian families of the 9th century derived their names from an ancestor, their successors named themselves after the area of their rule or their principal stronghold.

Entry into this later aristocracy was possible by a variety of routes. Throughout this period the enduring hazard of central authority was that those appointed to maintain its interest would succeed in converting this representation into a right. In Germany even the unfree class of *ministeriales* sometimes succeeded in establishing such claims to the imperial fiefs and castles entrusted to their care and thus forced their way into the aristocracy. In France and England in the later Middle Ages, royal servants were commonly enriched by the opportunity to deprive or buy out those outside the circle of royal favour. There, too, in the 14th century a certain number of great merchants in the service of the crown were able to enter the ranks of the nobility; in Italy, however, the incompatibility of merchant adventure with noble birth had broken down much earlier. In later medieval Europe the emergence of more professional armies (especially in Italy and France during the widespread disorders of the 14th century) provided opportunities for successful soldiers of modest birth to rise into the ranks of an aristocracy coloured by ideals of an elaborate knightly code.

Lesser nobility. It was of the essence of high social rank that the aristocrat should have a large following of men who were themselves of free birth. Every great man, like every king, was made great by his capacity to maintain a retinue and reward his followers. An estate was valuable less for its revenues in money and goods than for the men whose services could be secured from it. The noble retinue was held together by gifts of gold, horses, weapons, or hawks, while those who served their lord well might hope for a gift of land. The numbers of such followers, or knights, were swollen by freemen who were driven by need to surrender their own land to the lord and to receive it back as a conditional grant by him; in return they received a protection modelled upon that extended to the lord's own blood kin. With increasing specialization of warfare, free birth tended to give way to the skills of cavalry warfare as the essential qualification for noble service; these knights (French *chevalier,* German *Ritter*) might live as retainers in the lord's household or hold land from him—a knight's fee, or fief—coming only at an exceptional summons to his feasts, courts, or wars. Many enjoyed a large freedom in the government of their estates and came to form a lesser nobility, bearing arms, using their own seals, and living in more or less fortified manor houses. As the apparatus and conventions of knightly warfare became more elaborate and the influence of the courtly romances more pervasive, membership in the nobility came to depend on knighthood.

Knights

The rise of the knight to this more exalted status was in part a function of the dissolution of the tight bonds of dependence his land tenure would once have imposed. This was a general phenomenon of the later Middle Ages, for the complexity of tenurial obligations made them increasingly inadequate either for defining status or for providing the lord with his honorable retinue. The later medieval principalities therefore owed their legal constitution to the powers of the lord in the exercise of his sovereign rights rather than to his personal claims on the services of his vassals; by a parallel development, the lord's household and following were commonly then paid for their services. Even at the lower levels of society, the same movement can be seen at work. In the 9th century a great estate would normally have a proportion of household slave labour; by the 13th century almost all the household servants would receive some wages, though food, lodging, and security still formed the bulk of their payment.

The church hierarchy. The greater aristocracy throughout Europe included a number of churchmen, for the sees of bishops and many monasteries had received wide grants of land, over which they often exercised more extensive rights than did their lay fellows, both as lords

of tenants and as royal officers. Under Charlemagne and later princes, the state intervened to enforce universal acquiescence in the form of spiritual government over the whole body of the laity, and this could be burdensome or even oppressive. Set aside from their fellows to a different and a higher calling, this aristocracy and its subordinate officers were also, in theory, recruited on quite different principles. In practice, however, high office in the church remained almost wholly the prerogative of the aristocracy and later of the kinghtly classes. Between the 8th and the 15th centuries it has been calculated that not much more than one-third of the 2,000 bishops appointed to German bishoprics came from non-noble families, and only five are known to have come from the dependent peasantry that formed the great bulk of the population. Certain monasteries and colleges of cathedral canons were explicitly reserved to those of the most carefully authenticated noble birth. Yet access to the highest church offices was in part dictated by other considerations; from the 11th century onward, royal or papal service and mastery of the disciplines of theology and canon law provided the means for men to rise on their ability alone. Suger—abbot of Saint-Denis, a monk of modest origins who became chief adviser to Louis VI of France, and regent for Louis VII—and died one of the most powerful men of Europe, or Thomas Wolsey, a butcher's son who rose to be cardinal and archbishop of York and chief minister to Henry VIII of England, illustrate the advancement that the church could provide, however rarely. Opportunities for this upward mobility were closely related to the existence of an effective royal administration capable of excluding the local magnates from a monopoly of patronage; the hope of modest advances provided many obscure but devoted clerical servants for the infant bureaucracies of 12th-century kings.

Bishops

The original nucleus of church organization had lain in the bishoprics, groups of which, from the end of the 8th century at least, formed provinces normally presided over by an archbishop or metropolitan. The special interest of the church in political concord, which derived not merely from its dedication to charity but also from the vulnerability and wide extent of its property, made it for long the natural ally of kings and emperors; the anathemas of the bishops were regularly employed to reinforce the sword of the Lord's anointed. Being, if aristocratic, at least not hereditary magnates, the bishops were the allies upon whom the Saxon and Salian emperors of Germany and the early Capetian kings of France largely depended as a counterpoise to their greater secular subjects. Episcopal estates were treated as royal estates whenever the king journeyed about his kingdom, and the burdens of military service on German ecclesiastical estates were often unusually heavy; the royal administration was staffed and directed by the clergy, and the church's endowments were used to reward them. In return the bishops secured great privileges; the archbishops on either side of the Rhine—at Reims, Sens, Mainz, Trier, or Cologne—were at once powerful landowners and retainers of a wide measure of delegated royal authority, and these were only among the more visible examples of a movement in progress throughout Europe. Theoretically, the success of the Gregorian reform movement in the 11th and 12th centuries had stemmed from the distinction it drew between the bishop's absolute obedience to canon law and his subordinate obligation to the prince. But in practice, the bishops continued to depend on royal authority for defense against grasping lay neighbours, for the enforcement of ecclesiastical discipline (particularly in matters of heresy), and, in the later Middle Ages, for protection against the growing anti-clericalism of some of the educated laity and even the financial demands of the papacy.

Secular clergy

From the 5th century onward, the original constitution of the diocese, which rested upon the bishop and a small group of clergy living with him at a central point, altered substantially. Subordinate centres grew up within the diocese, some large and early ones served by groups of resident clergy but the great majority caring for only a small parish and served by a single priest. Recruitment to these was wide; some houses of canons and wealthy parishes provided revenues for men of high birth or influence, while others were served by men drawn from the peasant population, who were little more than domestic servants of the landowner who had built the church. The widespread grant of parish churches to monasteries and religious communities (as many as one-third of the parishes of a diocese could be so granted) and the prevalence of absentee clergy, such as pluralists holding several benefices or scholars engaged in study at the universities, meant that many churches were served by substitutes—vicars, who enjoyed only a proportion of the revenues of their office, often a quite inadequate one. Besides these, most parishes also provided some employment for a floating population of clerks in minor orders. With a rapid decline in the creation of new parishes in western Europe after the 12th century, it was the endowment of chantries, the setting aside of money and buildings exclusively for the saying of mass for the soul of the founder, that provided the greater number of new, if modest, benefices, though these were often attached to other parochial or charitable obligations.

The determined efforts made first by the early Carolingians and then by the reformers of the 12th and 13th centuries to provide an efficient system of supervision and control over the local pastoral work of the diocese ultimately produced an administrative hierarchy parallel to, though partly distinct from, the pastoral one. The rise of the bishop's formal jurisdiction saw the appearance of a host of legal and financial agents, ranging from the powerful archdeacon to the humble summoner, whose task was to enforce attendance at the church courts. The increasing centralization of church government at Rome in the 13th and 14th centuries required the appearance of other officers, proctors of bishops and abbots at Rome, collectors of papal taxes, and, later, such disreputable figures as the itinerant peddlers in papal indulgences and dubious relics—the pardoners.

Administrative hierarchy

Regular clergy

By the mid-13th century there existed beside this hierarchy another one, the regular clergy, living under a formal and corporate rule more demanding than the minimum canonical requirements of the seculars. Between the 5th and the 11th centuries this sector had been overwhelmingly monastic, composed of autonomous houses ruled over by an abbot with near absolute powers and devoted to the maintenance of the regular liturgical cycle; there was one pre-eminent rule, that of St. Benedict, which envisaged a life of corporate self-sufficiency as the norm. In the 11th and 12th centuries a number of variations on this rule appeared, designed to heighten the emphasis upon manual labour, corporate poverty, or solitary contemplation; in a more striking departure, some houses of canons under the flexible rule of St. Augustine were, it was hoped, to combine monastic rigour of life with active parochial work, while military orders founded in the course of the Crusades in the Holy Land and Spain and on Germany's eastern frontier combined the yet more disparate profession of war with monastic ideals. All these were property-owning corporations, as were the houses of nuns and canonesses that lived under similar conditions; their presidents, especially the abbots of the more well-to-do Benedictine abbeys, were often wealthier than many bishops and were drawn from a similar social background. Only among the Cistercian lay brothers was entry into the monastic life readily open to men of humble origin. The original followers of St. Francis (died 1226) were often drawn from a much wider variety of classes, but by the end of the 13th century both they and the followers of St. Dominic, originally intended to be educated preachers trained to combat heresy, had become orders of friars with a strong academic tradition, excluding all but the most talented of the very poor. These men renounced all property (even corporate), were especially engaged in the work of preaching (notably in the towns), and travelled constantly. By 1300 the multiplication of religious orders had come to a virtual standstill.

Other social groups. *Townspeople.* The urban society of the Roman Empire in the West was breaking down long before the deposition of the last emperor in the late 5th century, and it continued to fade with few interruptions until the 10th century. The estimated population of

Rome fell from more than 1,000,000 in the 1st century AD to 40,000 in the 7th; nevertheless, it long remained the largest city of the Christian West, although tiny compared to the greater Muslim cities or Constantinople. Many Roman towns continued to be occupied, often because they remained, or became, the seats of bishoprics or abbeys and so centres at least of consumption, or because their walls offered some shelter from a long series of invaders. Commercial activity continued in those ports of Italy still in touch with Byzantium and perhaps on a modest scale in some of the Lombard towns. It even increased in such northern ports as Duurstede in the Rhine Delta, the Viking entrepôt at Hedeby (in Schleswig), and the Russian cities trading through the Black Sea. By the end of the 11th century the recovery of urban life was more general. Merchants trading over long distances began to appear as an urban aristocracy, wealthy men anxious to secure a larger measure of control in the government of their towns. In Venice this development was already clearly visible, and the fleets and trade of this port stimulated a revival of town life throughout northern Italy. The origins of these early capitalists are much disputed: some may well have been fortunate peddlers, but more can be shown to have been members of the lesser aristocracy, or tenants or officers of great churches, with capital to invest in goods and transport. In Italy the participation of the landed aristocracy in the life and trade of the towns was shown by the presence of clusters of their tall stone towers within the city walls rather than on outlying hill tops.

Recovery of urban life

Increasing political stability on land and contact with the Levant by sea saw the rapid increase of such trading communities in size and influence. The Italian cities of Genoa, Pisa, Lucca, and Siena rose to challenge the earlier dominance of Venice and Milan. In Germany the Rhineland towns were already a political force to be reckoned with in the civil wars at the beginning of the 12th century, but they came to be overshadowed by the great prosperity and activity of the Baltic towns of the Hanseatic League, such as Lübeck, Hamburg, and later Danzig (Gdańsk), and later by the prosperity of such South German towns as Augsburg and Nürnberg. In Flanders textile manufacture, based in part on imported wool, produced in the 12th century a precocious industrial community with a large unstable proletariat concentrated in such towns as Ypres, Ghent, Bruges, and Arras; alone of the northern cities these could compete in influence with the greater Italian centres such as the manufacturing town of Florence (with a population of perhaps as high as 200,000 in the 13th century, largely supported by its textile industry) or the equally large Milan, celebrated for its metalwork and, most notably, its armories. Paris by the end of the 12th century was already acquiring most of the qualities of a capital city for the fast-growing area of Capetian power, with a population estimated at 80,000.

By the 13th century the Italian towns commonly contained a group of aristocrats by birth or by wealth long possessed whose political dominance was challenged by the more substantial of the merchants. Below these were the retailers and masters of the smaller crafts and then the wage labourers, apprentices, and beggars, who were normally without a political voice but whose grievances sometimes broke out in violence and even bloodshed. The urban nobility of Italy was without true parallels north of the Alps and it was the greater merchants, or merchant adventurers of the later phrase, who formed the directing elite in most towns. Although even at the end of the medieval period the towns' share in the wealth and population of Europe was still very limited (over most of the Continent perhaps not above one-tenth of the total), their influence was much greater. Since the 14th century, access to the towns had been an important agent in social change, and the continuing or growing prosperity of some 15th-century towns was in marked contrast to the widespread decline of agricultural production and profit. Politically, the merchant interests of the towns were of the first importance, for it was only they who could provide the large sums of ready cash with which the kings of the later 15th century established royal authority over the magnates. To the degree that the king could continue to enjoy the taxes of the burgesses of his towns, he could maintain or extend his authority; without them he was reduced to competing with his own magnates on little better than equal terms.

Agricultural society. The countryside during the Roman period was chiefly cultivated either by the slaves of the great *villa* estates or by more or less free cultivators, sometimes bound by the government to remain on their land in order to maintain a tax-paying population but otherwise not labouring under grave personal disabilities before the law. In the Germanic societies of the invasion period, a threefold division of the agricultural population was common—the people *par excellence* (the *karl, ceorl,* or *bôndhi*), free peasants cultivating their own land, bearing arms, and so participating in the public assemblies of *gau* or hundred; a more obscure group often described as freedmen (*aldiones* for the Lombards and Bavarians, *Leti* among the Germans, Franks, Frisians, and also perhaps in Kent), possibly freed slaves, possibly the survivors of an earlier conquered population who enjoyed limited rights but did not usually play an active part in the courts; and the thralls, slaves either captured in war, condemned to their state by the law, or reduced to it by penury. Such bondsmen might sometimes have their own huts, plots of land, and houses and therefore enjoy a minimal independence.

Germanic division of agricultural society

These forms tended to merge so that over much of western Europe by AD 1000 the characteristic villager (villein) held a substantial plot in the village fields but equally was expected to perform such onerous services on a lord's domain as ploughing, reaping, and carting and was subject to his lord's will in much the same fashion as the earlier landless slave. In addition, the village community would usually contain other men with smaller plots of land (cottagers), whose rents in labour and goods were correspondingly lighter, and also a small number of slaves, in the old sense of the word. All these unfree cultivators could be described as serfs, however diverse their economic conditions. In parts of Europe, notably in northern Germany and eastern England, the earlier free peasantry still formed an important element in the population, and by the 14th century their ranks were swollen by numbers of villeins whom changing economic conditions had allowed to commute their servile obligations for fixed rents in money. The settlers who had been persuaded to take up holdings in the planned land clearances of the 12th and 13th centuries by offers of considerable freedom of tenure had reached the same goal by different means. At the other end of the social scale, the landless wage labourer was now much more common.

The timing of this process, whereby many of the earlier free peasants were first assimilated to the legal if not the economic conditions of Roman slaves and then increasingly recovered their personal liberty to become rent-paying tenants, was extremely uneven. The cycle was complete for much of Italy by 1200 and for most of France and England by 1400, but the free peasantry of Scandinavia, parts of eastern Europe, and Castile was only beginning to feel the presence of a serf-owning class of landlords in the 14th and 15th centuries, at a time when it was fast becoming obsolete elsewhere.

Beggars and outlaws. A society that was at best only just above subsistence level maintained a large body of vagrants. Economic disasters such as plague or famine forced a desperate population to attack the crops of more fortunate neighbours; in the late 12th century, freelance mercenaries roamed Europe in search of employers and booty, while the 14th and 15th centuries saw more organized (and so more formidable) free companies of professional troops whose favoured fields of operation were France and Italy, where their captains traded their services like sovereign princes. All large towns either generated or attracted their share of beggars and petty thieves; in François Villon, 15th-century Paris produced a universal poet to express the pleasures and much more frequent miseries of this vagrant life. Town and country alike lacked any effective police force, and therefore, although the majority of known or suspected criminals could be dispossessed of land and property, even murderers were rarely apprehended. At times in 13th-century England only one in

The vagrant life

100 murderers was ever brought to trial and convicted. The rest fled, many apparently into the forest to live like beasts or find an organized band of outlaws, idealized tales of which became widely current in the later Middle Ages in the legends of Robin Hood. The local community protected itself as best it could against such bands, but the assistance of a local knight at the head of his retainers could rarely be distinguished from the nuisance it was supposed to abate. The fate of these outlaws was bound up with that of the forests, and both were in decline by 1500. But with the first signs of a recovery of population, vagrancy and unemployment were probably increasing.

Women. A society directed by warriors and celibate clergy was not one in which women would exercise extended rights very often. After *c.* 1100 patrilineal descent was almost exclusively the test of nobility, while matrilineal descent was often the test of serfdom. The property rights of women, though protected by canon and secular law, were confined; it was a cherished freedom for widows to be allowed to refuse a second match proposed by lord or kin. Practice, however, did not altogether conform to this appearance. In barbarian society in the period after the invasions, and sometimes long after, matrilineal descent was often important—the house of Charlemagne traced its origins back to a daughter of Arnulf, bishop of Metz. Although few queens ruled in their own right, many exercised great political authority, as in the minority of their sons; the regency of Blanche of Castile for St. Louis in France was a notable and successful example. The remarkable Queen Margaret succeeded in uniting the three kingdoms of Sweden, Denmark, and Norway under her regency in the Union of Kalmar of 1397, an act that influenced the future of all Scandinavia. Similarly, although debarred from the priesthood by their sex, a number of women played a leading role in ecclesiastical affairs. St. Catherine of Siena and St. Bridget of Sweden played a major role in achieving the return of the papacy from Avignon to Rome in 1377; both were celebrated adepts of the spiritual life, the female contribution to which is also attested by the works of Margery Kempe in England. The influence of women was also felt in their role as patrons, sometimes of Christianity itself, as when a number of the invading chieftains of the 5th and 6th centuries were first married to Christian princesses and then converted. Queen Margaret of Scotland (died 1093) was responsible for the thoroughgoing reform of the church in her kingdom, which brought it rapidly into the mainstream of the Latin Church; another notable lady, Matilda of Tuscany (died 1115), had been the last refuge of the reforming papacy for almost a generation.

As patrons of the arts, and above all of poetry, Eleanor, duchess of Aquitaine and wife first of Louis VII of France and then of Henry II of England, and her daughter Marie, countess of Champagne, were the leading figures of their century. At Marie's request, Chrétien de Troyes translated Ovid's *Art of Love* and also wrote one of the first courtly romances to be based on Geoffrey of Monmouth's Arthurian history. In Marie de France the courtly romance found a skillful female writer—appropriately, because these tales, which evolved from the masculine world of the chanson de geste, reveal the growth of a new social convention in which women had a larger part and a higher function. Foreshadowed in the diffusion of the cult of the Virgin Mary in the 11th century, this new convention was developed in the Italy of Petrarch and Dante.

Medieval economic patterns. Latin Europe by 1500 covered a great diversity of lands and climates. A first division was that between the heavy soils of Germany, northern France, and Britain and the predominantly lighter soils of the Mediterranean lands of Spain, Languedoc, and Italy, in which vines and olives formed an important adjunct to grain.

Medieval communities. Both main types of agriculture were conducted by cultivators who lived, where possible, in large settlements, with their churches, mills, and barns, the whole settlement often walled, as in the *bastides* of southern France, or at least hedged against animal and human marauders. Both economies contrast with those that existed on their perimeters and in the less fertile or more broken country of the interior. In the Celtic lands of Ireland, Scotland, and Wales, in Spain and the foothills of the Alps, as well as in Hungary and the Latin Balkans, a largely pastoral economy flourished, depending on flocks of cattle, sheep, or goats and producing a quite different pattern of living. As oxen provided the essential power for the cereal farmer's plough, and milk, cheese, salt, and meat were as important to his diet as a minimum of crops were to the most pastoral of societies, so most medieval communities represented a precarious balance of forces between the needs of crops and beasts. Beyond these types of communities there existed various types of highly specialized settlements, such as the fishing communities of the North Sea coast, the salt evaporators of southwestern France, the fenmen and the miners, sometimes solitary and sometimes organized in tightly knit communities, as were the tin miners of Cornwall and Bohemia. With few technological resources and poor communications, all medieval communities were largely conditioned by their environment.

Some of the Mediterranean communities of Italian cultivators had been organized as great estates (*latifundia*) in the late and post-imperial period; each tenant held only as much land as could support him and his family while he worked on the lord's estate. Very early, however, the comparatively widespread use of coinage allowed landowners to abandon such direct exploitation, which was clumsy and time consuming, in favour of a variety of leases for terms of years or lives. By such means the community became one of independent cultivators, each with his own field. Similar tenancies prevailed over much of southern France, where a Roman Law much modified by custom and the decay of the judicature continued to regulate contracts and the functions of the market.

The triangle formed by the rivers Loire and Rhine contained the chief area of nucleated cereal-growing settlements, though there were notable extensions of this into Britain to the west and into Franconia, Swabia, Bavaria, and the German east. In the Loire–Rhine region, with its heavy soils and wet climate, the earlier small enclosed fields of the Iron Age gave way widely to the characteristic open fields of the medieval vill, where the arable land of the community lay in large blocks cultivated by heavy, ox-drawn ploughs. The mechanical difficulties of turning such an implement and the need for effective drainage produced a characteristic effect of long narrow parallel strips running down contours of the hills. Such cultivation required a regular alternation of crops because there were few means available for restoring the fertility of the soil except allowing either half or one-third of the land to lie fallow each year. Similarly, the other resources of the land in pasture, woodland, water, and common grazing needed careful annual regulation because there was constant tension between the needs of men, crops, and beasts. This required an organism whose common life maintained some continuity regardless of the divisions that accidents of lordship might impose; later, the village was often a legislative body, enacting its own bylaws.

Open fields of the medieval vill

Until the 13th century this open-field cultivation continued to support the classical manorial organization. The lord drew his profit partly from his possession of a share of the vill's arable land, which was exploited for him by his tenants for some of the week, and partly from the exercise of his right to compel his tenants to use his ovens and mills, at a price, and to control and exploit the use of the surrounding waste land. In practice such communities had always maintained a population of landless men—slaves earlier and labourers later—and there usually were tenants who owed rent or personal services of a less strictly economic variety. (Such free tenancy became widespread with the clearing of outlying lands in the 12th and 13th centuries, especially in the planned colonies along the frontiers.) Again, principally under the impact of increasing internal trade and a greater circulation of money, cumbrous forms of exploitation tended to break down in the later Middle Ages as the domain was leased out and the lord allowed his tenants to commute their labour dues into cash rents. When, in the late 14th and 15th centuries, the relation between land and labour shifted

drastically in favour of the tenant over most of western Europe, the economic centralization of the local community focussed on the lord's estate almost vanished, though the need for common action in the agricultural cycle did not. In contrast to these tightly knit, often productive and vulnerable agricultural communities stood the much looser society of the higher lands, where settlement was necessarily dispersed and where social bonds rested much more exclusively upon the ties of the kin or clan than they did in the nucleated villages. Even the arable farmers in a scattered community had more independence than did their lowland fellows. The Icelandic sagas written down in the 13th century, in manuscripts typically preserved in farmhouses rather than in monastic libraries, provide a vivid picture of such communities.

Increasing volume of economic exchange

The increasing economic specialization of the period between 1100 and 1500, which saw vine growing and sheep rearing rise to the status of basic rather than supplementary occupations for whole communities, was made possible by a great increase in the volume of exchanges, itself a function of greater political stability. These exchanges took place pre-eminently in three settings: the markets, chiefly for local produce, held at short and regular intervals; the trading communities of the towns; and the occasional great fairs, annual events to which merchants travelled from the ends of Europe and the proceeds of which enriched such fortunate princes as the count of Champagne, lord of the fair towns of Provins, Troyes, Lagny, and Bar-sur-Aube.

Urban growth. Early concentrations of population in settlements occurred for political, ecclesiastical, or defensive reasons as often as for commercial ones; but the period from the 11th century onward saw the widespread rise of classes of men engaged in the commerce of exchange or manufacture and settled for that purpose in urban groups. Many of these were originally only merchants in a subsidiary sense; and many towns were barely distinct from large villages, being surrounded by the town fields cultivated by the citizens and containing within their boundaries numerous livestock and small patches of cultivated ground. Yet, a common interest among merchants or craftsmen produced associations of considerable importance. Merchant guilds developed—groups of men whose interests extended far beyond the political horizons of the lords of most towns, anxious for the reduction of tolls and in need of their own courts of justice appropriate to the urgent demands of itinerant commerce. It was these who led the struggle for urban autonomy in Europe north of the Alps—a movement which almost always involved conflict with the local lords, especially bishops and abbots, but was often favoured by princes such as the counts of Flanders or Henry the Lion, duke of Saxony, in the German east—and thus the juridical phenomenon of the town as a corporate person emerged, created within a pattern of territorial lordships. Similarly, such towns acted corporately, much like their contemporary landed magnates, seeking to subordinate neighbouring communities to the authority of the city magistrates, to compel the man of the countryside to acknowledge the legal and economic primacy of the urban centre, and to divert profitable trade routes to their own advantage. The 12th-century crusading movement reflected such local hostilities—the Byzantine monopoly enjoyed by Venice was challenged by the direct access to the trade of the East offered by the crusader states, often maintained by the rival fleets of Pisa, Genoa, or Lucca. Long intercommunal wars, such as those between Pisa and Genoa or between Venice and Ravenna and Pola (now in Yugoslavia), saw the rise of a small number of city republics governed by elected consuls combining birth and commercial substance in a unique blend. Within such towns there existed other interests also driven by their common economic interests into corporate action.

In some towns, such as London or Bologna, a civil organization based on a geographical division into wards or quarters existed in potential conflict with the alternative divisions by occupation or wealth. Associations of craftsmen and retailers formed themselves in guilds that had many objects beyond the commercial. Among the earliest known forms of such guilds were those of prayer; and many guilds had their own chaplains, churches, and such distinctive religious pageants as the cycles of plays performed by the various mysteries (or crafts) at Wakefield in England. Many also maintained a form of insurance for their members, for the old, and for the widows. But all came to be devoted largely to the regulation of their crafts—especially entry to them—and to the struggle for their interests against the competing ones of their suppliers or of related trades. The interests of the masters of these crafts were often opposed to those of the greater merchants, thus producing a struggle for control of the machinery of town government. Equally, however, other divisions cut across this classification by trades. Within the guilds there was a struggle between masters and men, between those anxious to reduce competition and those bent on a larger independence. Occasionally, and most strikingly in the industrial towns of Flanders in the 13th and 14th centuries, this produced proletarian risings of wide impact; at Courtrai in 1302 an army largely drawn from the Blue-Nails, the hired dye workers, inflicted a defeat upon the forces of the King of France and the Count of Flanders; frequently, the larger Flemish towns acted to free themselves from the power of the count or to maintain their communications with the English wool trade against the pressures of France.

Struggles for the control of town government

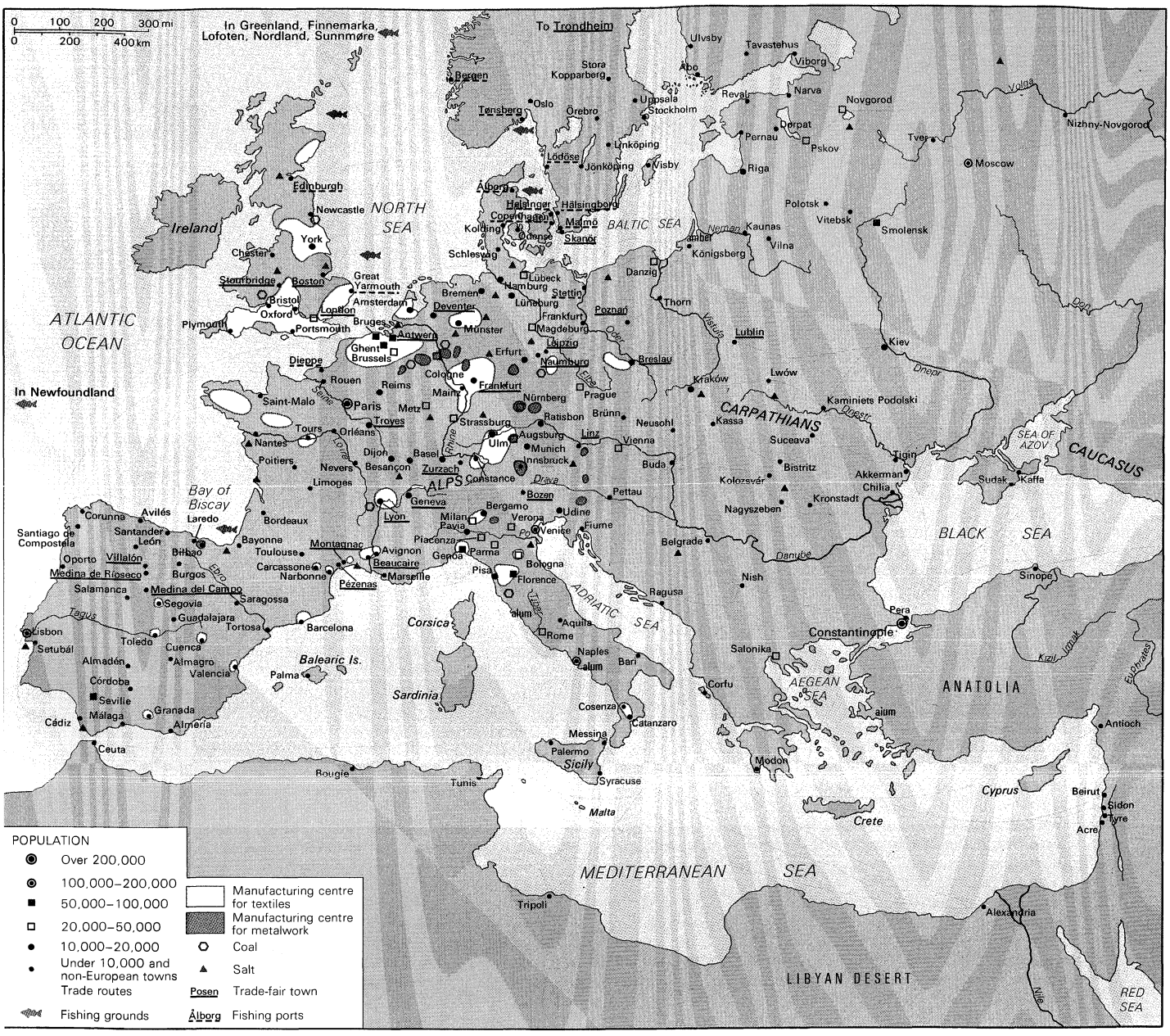
Though an urban proletariat of this kind was rare in medieval Europe, conflict between an urban oligarchy of landowners or great merchants (the *popolo grasso* of Italy) and the lesser craftsmen and labourers (the *popolo minuto*) was extremely common; it was at its fiercest where the cities enjoyed great autonomy, especially in Italy, where there was no effective or extensive central authority to maintain order. Hence, there were frequent civil wars in the Lombard towns, complicated by their external relations with each other, with the papacy, and with the empire; and there was a wide variety of forms of communal government, ranging from such attempts at radical democracy as the rising of the Ciompi of Florence in 1378 to an increasingly dominant model of dictatorship, such as that exercised by the Malatesta of Rimini, the Visconti of Milan, or even the Medici of Florence.

The independence of the Italian city-states, which was complete by the end of the 14th century, had rested in part upon the successes of the Lombard League against the empire in the late 12th century. North of the Alps, particularly in Germany and the German-settled lands of the east, towns established similar but rather more restricted liberties against the greater princes and formed into leagues of economic communal interest, of which the most notable was the Hanseatic League of Baltic towns, which secured a near monopoly of the profitable northern trade in timber, furs, wax, and amber. First clearly organized in the late 13th century, by the 14th the league was the dominant commercial force in northern Europe. In alliance with the Teutonic Knights and by an aggressive policy of embargo and blockade, it secured privileges superior even to those of the native merchants in Russia, Sweden, Norway, England, and Flanders. Only in the later 15th century did the league begin to decline, with the failure of the Knights, the increasing commercial activity of the western kingdoms, and the decline of the Baltic herring fisheries.

Urban independence on the Lombard or Hanseatic model was exceptional, however; in France, Spain, parts of Germany, and England the towns remained firmly within the framework of royal or princely government, not least because only the king could be relied upon to maintain political and fiscal stability at home and to negotiate commercial privileges abroad. Contingents from the towns formed a numerically formidable element in the royal army at Bouvines (1214), when Philip Augustus defeated a combination of Norman, Flemish, and imperial forces aiming to restore the earlier independence of some of the greater vassals. The protection of merchants had been a duty to which 11th-century popes had recalled kings; it was the financial support of the towns that was of importance in the triumph of the "new monarchies" of 15th-century France, Spain, and Tudor England.

Protection of merchants

Forms of lordship. Upon these economic structures there rested a variety of structures of lordship. The earliest



Medieval industry and trade *c.* 1500.

From *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

that can be discerned in the centuries after the migrations were extremely complex, being functions of a power that was at once one of property, of kinship, of public function, and even sometimes of priesthood; only in the course of the later medieval period did these elements of authority come to be distinguished and sometimes divided.

Serfs. The economic bases of lordship in its least complex form lay in the ownership of men: the large household of slaves was the lord's property, and they worked his land. They enjoyed no rights against their lord, and he was responsible for maintaining peace and regulating their duties; such households of slaves became rare quite early over most of Europe. Domestic slavery was still an active force in the Iberian peninsula and Southern Italy at the end of the Middle Ages, though only in Sicily and the Balearic Islands were slaves widely engaged in large-scale agriculture. The responsibility of the lord for his own domestic servants remained near absolute, the successors of many household slaves were the serfs who settled on small plots of land; the descendants of other freer men were reduced to a similar status by poverty, war, the burdens of public taxation or even by the justice of the church. Like the Roman slave, the medieval serf had no public rights against his lord; he was *adscriptus glebae*—so bound to his land that he could not leave it, and a part of its stock to be sold with it. The lord, by virtue of holding the land, was responsible for all police jurisdiction over his unfree men in his own court. His serfs were unable to marry without the lord's consent; in theory, their goods were his to tax at will during their lifetime and to confiscate at their death; their children were born into serfdom; and in all disputes there was no appeal from the lord to any higher tribunal.

Practice diverged at least in part from this grim theory. The agricultural cycle in which these serfs lived was conservative and complex, and custom operated powerfully to maintain it. Correspondingly, the disputes of tenants with each other and with the lord were regulated essentially by local custom, which was proclaimed by the body of the tenants in the lord's court. The rights of the lord to the labour of his serfs through the year and at the chief seasons of ploughing and harvest were also fixed by custom and were rarely (and slowly) altered except at times of sudden crisis. The rights of the lord to the property of the serf were equally confined, so that the serfs' duties were expressed as the obligation to render produce for the great

feasts of the year or to take their grain to the lord's mill, their flour to his ovens, or their grapes to his wine press or to make fixed payments at the marriage of a daughter, a father's death, or entry into a peasant holding, whether by inheritance, marriage, or purchase. Any or all of the other services could be commuted for money payments also, and, in western Europe at least, the function of the lord's court in regulating serf labour on his demesne was becoming obsolete by 1500, though its other functions still had a long future.

The bond between serf and lord

In some parts of Europe the duties of the serfs included services that are less comprehensible; and the very intimate bond between serf and lord might contain elements more accessible to an anthropologist than to an economist or historian. In many cases the autocratic power of the lord was mitigated not merely by force of custom but also by close and frequent personal contact. The plans of early manor houses and castles show no extensive private rooms for the lord's use: he dined with his servants in his great hall, where he meted out justice. There was, until the later Middle Ages, little provision for private gardens for the lord's use; all but very great men could be found working at their own harvest. And feasts in the hall brought together the whole village community.

The lord had need of other services beyond those of his house and arable land; the greater Carolingian estates had even carried their complement of carpenters, smiths, potters, and weavers, free or unfree, as well as the usual haywards, shepherds, and beekeepers. Men personally as unfree as the serf might also undertake more responsible tasks; the *ministeriales,* pre-eminently of Germany, acted not only as bailiffs but even bore arms as knights, though in such cases the dignity of the occupation came ultimately to cancel the defect of birth.

Feudal bonds. Men of free birth were found over all Europe, though in varying numbers, who recognized the authority of a lord for some purposes—holding his land in return for rent or services less servile than those of the villein. Most areas of Scandinavian settlement, much of Saxony, and the Low Countries were marked by numbers of such men, who often pronounced judgments in their lord's court or escorted him through his estates, providing a contingent of men-at-arms in war and rent-paying tenants in peace. Lordship of this kind stemmed directly from the ownership both of land and of some of its tenants; the profits of lordship were, however, drawn from much wider sources than the labour of serfs or the rents of free tenants. Either by grant or by usurpation, a great variety of forms of indirect taxation were open to the lord. The levy of tolls on rivers, bridges, or roads was perhaps the easiest source of revenue open to any man; a license to hold a market in a vill would not be sufficient in itself but also involved payments for the holding of stalls and even a tax on individual transactions. Very great men might even mint their own coinage; consequently, the currencies of France, Germany, and Italy were extremely fragmented and often unstable.

The most distinctive form of medieval lordship, however, was that which is often called feudal, from the Latin *feudum,* meaning "fief," which was its central feature. In essence this was a fusion of the earlier *precaria,* a grant of land made for a fixed term in exchange for services or rents, with the very general commendation by which a man placed himself under a lord's protection by becoming his man or vassal. Some of those who had served their lord well would receive from him a benefice of land or revenue as a reward; others in dire need of protection would surrender their own land to the lord to receive it back as a benefice from him. When the tenant's right to his land became hereditary and his tenancy, or fief, ceased to be a reward for past services and became the reason for services to be performed in the future, feudal tenure was fully developed. These processes can already be traced before the end of the 9th century in the Carolingian Empire.

Creation of feudal bond

By the 12th century such tenure was to be found throughout Latin Europe. It was characterized by a number of symbolic acts. The first was homage, the process by which the man knelt and placed his hands between those of his lord, so putting himself at the lord's disposal and under his protection. The next was the oath of fidelity, sworn by the man to his lord, sometimes sealed with a kiss. Then came investiture, by which the lord handed over some token of the fief to his new man. This sequence was first fully described in the year 1127, but its various elements can be traced or inferred very much earlier, though they were not all necessarily found together. The bond so created was much more than a form of land tenure; it was first a human relationship, in which the lord assumed many of the rights and duties of a father and from which the man could escape only if the lord directly attacked his life or family. If the lord died leaving a child as heir, it was the duty of the tenants to maintain the heir's rights until he came of age; similarly, if a tenant died leaving children under age, their wardship was the lord's. Not the least of the man's obligations might be that of attending the lord at the great feasts of the year, which were at once parties, parliaments, and law courts.

The duties of the lord to the tenant were usually only generally stated; he was bound to protect his man in war and peace, in the field, and in the law court. Carolingian legislation sought to ensure that every man had a lord to answer for him; a lordless man was a man in danger himself and a danger to others. The duties of the man to his lord varied enormously. Sometimes, even very early, the terms of his tenure were minutely defined. More often the general phrases of aid and counsel were called on to cover every possibility. As in the case of the servile tenant of the vill however, a general subordination to the lord first became fixed by custom and then often commuted into a money rent. The aid the man owed was primarily military aid, essentially service with the full equipment of a knight—lance, sword, helmet, mail hauberk, and powerful horse. A great man's household usually contained a permanent body of these knights, who were often landless young men of good birth but no fortune; but the obligations of the enfeoffed knights living in the estates they held of their lords were early restrained, by custom at least, to a period of service in the field of 40–60 days and sometimes a limited period of garrison duty or castle guard at the lord's fortress. Not all personal aid, however, was military: there were also the sergeanties—fiefs held on condition of performing other services. These ranged from fulfilling the highest offices in the lord's household, such as steward, constable, or marshal (or even jester), to picturesque or purely honorific obligations, such as providing the lord with an annual goshawk, a leash of hounds, or a pair of gloves. All such services might come to be commuted for money payments; by 1200 it was common for the bulk of the knights liable for royal and even magnate service to settle their obligations by the payment of sums of scutage (literally shield money). Because, by then, service was seen as a burden on land and because this land might be in the hands of a church, a minor, or a woman or might be divided up among many heirs, such payments were a convenience to the tenant. The campaigning season and the high costs of a changing pattern of warfare often made money payment just as attractive to the lord.

Payment of scutage

Sometimes the lord might need financial aid for less strictly military reasons; again, custom came to distinguish the aids that he could levy from his tenants by right from those extraordinary ones for which the tenant's consent had to be sought. Practice varied, but the four most frequent were the knighting of his eldest son, the marriage of his eldest daughter, his setting out on crusade, or the ransom of his body.

Although in all these cases custom tended to establish the tenants' right in their land, there were three important traces of the lord's continued interest in what his ancestors had once granted: the relief, payable by an heir on entering into his inheritance; escheat, by which the lord recovered control of a tenure for which no direct heir could be found; and forfeit in the case of a tenant's treason or failure to perform his duties. The financial exploitation of these incidents of tenure and of the rights of wardship and marriage was among the most widespread of grievances against kings and great magnates in the 13th and 14th centuries.

The counsel that a tenant owed his lord also acquired a

formal sense. By virtue of having tenants, a lord had an honour court, where disputes between tenants or between lord and tenant were heard. The tenant's duty to attend (to perform his suit of court) was of the greatest importance to the lord: it was the number and dignity of the suitors to his court that gave its judgments authority and stability. Around this central institution of the lord's court grew up in the greater honours much of the apparatus of a sovereign state. The greater officers of the lord's household, most notably the steward, or seneschal, played an active part in overseeing the lord's estates, and in the later Middle Ages they often formed a council with regular sessions to audit the accounts of their lord's estates and developed their own code of legal precedents. In the 13th century, treatises upon the customs of such tenurial courts appeared in considerable numbers side by side with studies in Roman or royal government (the *Sachsenspiegel* in Germany or the work of Beaumanoir in France may serve as examples). Only their homage to the king or emperor, with a variable liability to be summoned to his court, distinguished the greater magnates of France or Germany from sovereign princes.

Hierarchy of lordships

Theoretically, society could be conceived as a hierarchy of such lordships, with the sovereign at its head; but the practice was usually very different. In Germany the obligations of vassalage long retained traces of their servile origins, so that great men assumed them only with reluctance. It was correspondingly rare for a man to be the vassal of more than one lord; hence, a hierarchy of homages could appear in the late 12th-century *Heerschild* (a formal definition of social standing according to the number of intermediate lords between a man and the emperor) with the emperor at its head, the greater churches beneath him, then the imperial princes, who had done homage only to the emperor or the church, then counts, then noblemen, and so on. Even here, however, the consent of the other princes (the *Reichsfürstenstand*) was necessary for admission to the highest ranks, and the obligations such homage entailed were relatively very slight. Elsewhere—in France particularly—homage to more than one lord was frequent (the county of Champagne was held of 10 different lords), so that no such organizing principle could operate. What determined the permanence and vitality of tenurial networks was in part political and military circumstance, in part the extent to which some higher authority was able to intervene between lord and vassal, and in part the extent to which this lordship over men or tenants was able to absorb earlier administrative or public authority.

Royal government. The growth of the centralized institutions of predominantly royal government is a phenomenon that was involved with the rise of certain technical skills in the collecting and organizing of information. Until the 13th century there were no maps of great practical value; the statistics occasionally collected by medieval governments were often inconsistent and made according to erratic principles. Until near the end of the period, the usual methods of accounting and bookkeeping precluded most calculations performed by a modern government. It is a notorious defect of medieval chroniclers that the numbers they attribute to populations or even comparatively small armies are almost invariably the purest fantasy. The technology of transport was equally primitive; the best roads in the 15th century were still the Roman imperial streets, for all their thousand years of neglect, and winter flooding still affected most river systems in spite of some large-scale drainage works that had been undertaken in the Netherlands, Italy, and elsewhere. Under such conditions a central government wholly free of "constitutional" restraints would still be obliged to leave its local agents a large measure of independence, and the custom of the neighbourhood was necessarily the chief regulator of political as well as social and economic relations.

The royal household. The seed from which all medieval institutions of central administration were to grow was the immediate personal household of the king; its members were the only permanent staff he had, and only their intimate personal association with their lord was a guarantee to others that they did indeed represent his views and bear his authority. The essential elements of the early royal household therefore provided the framework of the first departments of state.

Officers of the hall

The chief elements of the royal household were four—the hall, the chamber, the chapel, and the courtyard, with its horses and stables. The whole household, but in particular the hall, which was at once palace, law court, and dining room, was directly governed by the steward (*dapifer, seneschal,* or *drost*), under whose direction the guests were seated and the feasting conducted, while the wine was under the charge of a butler (*pincerna, Oberschenk*). Under the later Merovingians in Gaul, the mayor of the palace (literally chief of the house) became a figure so powerful as first to overshadow and then replace his king; elsewhere, though less powerful, the steward was the usual chief deputy of the king. Under the Capetians of France he was charged with the annual scrutiny of the accounts tendered by the king's local bailiffs, the provosts; in Normandy, Jerusalem, and elsewhere he was the natural choice as regent in his lord's absence; the butler shared some of the prestige of the steward but had yet few defined functions.

The chamber

The chamber, the room in which the king slept or took private counsel, was also the natural place to store his treasure; hence, the chamberlains were often specially charged with the collection of revenue and handing it out as the king had need. The papal treasury was known as the Apostolic Chamber, and the papal chamberlains were widely entrusted with financial missions.

The chapel, containing the royal altar and relics (the term chapel derives from the *capella,* the short cloak of Saint Martin preserved among the chief relics of the Carolingian royal treasury), was served by chaplains, to say the daily mass, assisted by a body of clerical assistants. As the chief and sometimes only literate members of the household, these men were responsible for drawing up such documents as the earlier kings required; among their number and often at the head stood the chancellor, whose special task was the authentication of royal acts, usually with the seal, which he kept.

This royal household was constantly on the move, carried on carts or packhorses, and hence the great importance of the last two major household offices, those of the constable and marshal. The duty of the count of the stable (constable) was closely associated with the organization of the army, and hence the term came to be used of commanders of garrisons as well as the central household officer. The office of the marshal was originally more humble but shared the military fortunes of that of constable. In the 14th and 15th centuries constable and marshal came to a new importance as military commanders and correspondingly acquired judicial competence as presidents of the courts of chivalry, which dealt especially with military discipline, the division of ransoms, and the right to bear a coat of arms.

The growth of a permanent bureaucracy. These early household offices, with characteristically unspecialized duties, changed greatly under the impact of two forces. First, these central offices had a tendency to become hereditary, much as the local offices of count or duke had done; since the domestic service of the king, at least on public occasions, was itself a very great dignity, the most powerful families claimed the right to perform it, so that the office came to be the prerogative of great magnates who were too preoccupied elsewhere to perform their duties in person. Even the chancellorship sometimes became attached to certain archbishoprics—Reims or Mainz, for instance; since the king's domestic needs continued, a distinction evolved between the hereditary dignitaries such as high steward or archchancellor and the men of much humbler rank who actually performed the routine duties of the household. Second, the increasing volume of business done in the king's name—judicial, administrative, or financial—demanded ever more elaborate records and a more extensive staff; therefore, the offices of government were less mobile, and a physical distinction became common between the constantly itinerant household about the king's person and the more cumbrous (though rarely wholly static) departments of permanent officials. Furthermore, the processes of government, especially in the

chancery or (particularly well documented) the English exchequer, became arts or mysteries that demanded a staff of financial or legal experts with a specialized training. Thus were born the chief departments of state; well beyond the end of the medieval period, however, their principal officers were still considered the king's servants in a literal sense. This household character of public office made the distinction between loyalty to the king's person and loyalty to the office extremely hard to draw and frustrated many early efforts at "constitutionalism."

The chancery

Of these departments of state, the chancery was perhaps the most essential, for without a means of transmitting a number of recognizably authentic commands or recording the business already done, no large measure of centralized activity was possible. The first sign of the emergence of a chancery proper is the appearance of stereotyped formulas for the drafting of documents, classified according to their purpose. The papal chancery was the earliest office to develop this skill on a large scale; the rhythm of the text and the forms of authentication for papal bulls were already formalized before the mid-12th century. Similar tendencies are found in England in the reign of Henry II, and in France only a little later. (The urban notaries of Italy and parts of southern France were already using formalized business documents, though for a much more restricted area.) The second sign is the appearance of a substantial collection of records; apart from the financial records, dealt with below though often compiled by chancery clerks, the essential element was the keeping of copies of documents issuing from the chancery, distinguished according to their character. Papal registers had probably been kept from a very early date, and a later copy of the Register of Gregory I still survives; an imposing and ever more complex consecutive series of original registers survives from the early 13th century. These were in the form of books. In England, where the great period of initial expansion lay in the period between 1190 and 1216, and in France, where the early royal archives have suffered much graver losses at the hands of time, the characteristic record was a series of parchment membranes stitched together to form long rolls. These archives evolved rapidly through the 13th, 14th, and 15th centuries, with the preservation of many more classes of document, including informal memoranda. Such records served the purposes of both governor and governed, for they provided precedents and the material for the reform and improvement of administration for the king's servants. They also provided authentic copies of title deeds or privileges for a subject at odds with his fellow or even the king himself. The conventions of the chancery had a marked tendency toward autonomy; efforts at magnate control in England or extensive reform of papal administration in Rome were partly frustrated by the elaboration and conservatism of chancery procedures that might at other times offer a useful defense against arbitrary government.

Among the most immobile elements of medieval government were those connected with the collecting and checking of the royal revenue. Until the 13th century, the only coinage current was the silver penny (or denier), so that large sums could only be transported packed in barrels of great weight. It was therefore natural for royal treasuries to appear as places of permanent deposit, from which the itinerant court or army could be supplied at need. Often the treasury would also be the first site of any permanent archives. The rendering of accounts by the sheriffs, provosts, bailiffs, or seneschals who were local collectors of royal revenue was a matter that could not be performed at a wholly itinerant court. Not only was it necessary for auditors and agents to have a known place and time at which to meet, but all except the least sophisticated forms of accounting required the keeping of records of former debts and present liabilities. Under Charlemagne, a *Capitulare de villis* envisaged a wide enquiry into the estates of the emperor, though only fragments of the returns survive; similar surveys were being made in 11th-century France and Germany on a small scale, but no early text can rival the record of Domesday Book (compiled in 1086/87), a survey of almost all of England, shire by shire and fief by fief, drawn up by the clerks of William the Conqueror. It is still preserved among the English public records.

The growth of permanent records

The rise of the exchequer as a semipermanent office and court for the hearing of accounts and the adjudication of financial claims was an early and striking feature of Anglo-Norman government on either side of the Channel. The exchequer took its name from the checkered cloth used by the royal officials as an accounting device. The cloth was used as a kind of abacus and may have owed something to Arabic skill in mathematics transmitted by the Norman settlements of Sicily. Certainly in existence soon after 1100, its first surviving pipe roll of 1130 is the earliest record of receipts of its kind for all of medieval Europe. In France the Capetian kings for more than a century used the Knights Templar as their bankers, so that it was not until after the withdrawal of his treasure from their hands by Philip the Fair in 1295 that a fully independent Chambre des Comptes emerged, with a comprehensive staff of financial experts.

The cumbersome machinery of such financial bodies was a grave impediment in times of political, financial, or military crisis, although as in the case of the chanceries, due process offered some protection against arbitrary government. The frequent wars of the English kings and their long absences in France, as well as their desire to escape the oversight of great officers approved by their magnates, encouraged the appearance of various financial systems within the household to compete with that of the Exchequer; the wardrobe under Edward I, for example, expanded from its original function as the king's domestic treasury to become the organ responsible for the payment of his whole army or, under his descendants, of the chamber itself.

The king's council

As these departments of state became more fixed, professional, and independent, royal councillors who gave a political direction to the government acquired a distinct existence. From the beginning, the king's household had been of shifting composition, and a number of men enjoying his special confidence were often to be found there even if they were not holding household office. In the early 13th century the king's more or less permanent advisers began to take on a more formal character as the king's council. This council, often given definition by the taking of a common oath, was as omnicompetent as the king it served, and correspondingly, in times of political crisis, magnates or great assemblies sought to impose their nominees—as in England in the crisis of 1258–65 or France in 1356–57 or Aragon under Alfonso III. The direct and personal nature of the council's functions made such efforts almost always abortive. Only in Scandinavia, where the union of the three crowns involved prolonged regencies and uncertainties of succession, did the Råd enjoy a large independence from its king, though many German princes of the 14th and 15th centuries were compelled by their estates to accept a nominated council. In the 15th century the powers of last resort possessed by the king and exercised in consultation with his council allowed the formation of offshoots of this council as prerogative courts of justice, administration, and finance.

Birth of parliamentary bodies. Over the same period, the larger assemblies in which some royal action had long taken place were also taking on a clearer form and defined functions, issuing in the assemblies of estates to be found over most of Europe in the 15th century. Three distinct influences lay at the origin of this development—ancient custom, feudal law, and administrative convenience. The Germanic tribes described by the Roman historian Tacitus in the 1st century AD gathered regularly in arms to determine matters of general importance. In some respects it was the size of such gatherings that decided the size of the kingdoms to which they elected chieftains; such gatherings of the warriors continued well into the period of Frankish rule, even in those areas where the powers of the king were very great. In Scandinavia meetings of the *thing* never wholly passed into the power of magnates or royal officials, though elsewhere they changed radically.

As the old *mallus* or *thing,* with its lawgivers and its president, had become the court of an hereditary count and his vassals, so the assembly of freemen was widely altered

in the same way. To the king's great courts his chief tenants were summoned by virtue of their obligation to give their lord counsel; the number of those who came was the critical test of the king's authority. The decline of French royal power in the 10th and 11th centuries or of German kingship in the 11th and 12th centuries can be plotted on maps by examining the composition of their great courts, especially at the chief feasts of the Christian year (or at principal landmarks in the life of the royal family, such as marriages and the knighting of the eldest son).

Since these large sessions of the tenants were also the most solemn courts a king could hold, great issues would be brought to it; since they were usually held at known times and places, those who desired the king's help in securing justice would seek them out. There were, however, further reasons why the king should try to enlarge the attendance at his courts or councils; the bonds of land tenure in most of Europe before 1300 had long ceased to articulate all the elements of society. To regulate the affairs of merchants, for instance, the feudal bond was worthless to secure counsel or consent. More important still, the obligations of feudal tenure no longer provided the forms of taxation needed for the defense of the state, and whatever methods of raising money supplemented them required the consent of the community in a new sense. The princes of the late 13th century claimed to be lords of states, not merely of associations of men; in their conflicts with the papacy, in their assertion of legislative authority, in their claims to the financial support of the whole community, kings required a very general assent. Unless qualified representatives could be gathered, the king's will could not be known or the justification of changes publicized; the rise of the representative assembly is parallel to the rise of royal propaganda.

Representative institutions

In the 13th and 14th centuries these considerations produced a variety of experiments. In France after 1300, two meetings of the estates of magnates, churchmen, and burgesses were held often: one for the provinces of northern France (Languedoïl), another for culturally different provinces of the south (Languedoc). Provincial gatherings of this kind were naturally prevalent in the states of Germany, where the dynastic disputes and poverty of many of the princes made them peculiarly subject to the pressures of their estates of knights, burgesses, and clergy; the imperial Reichstag, attended theoretically by the tenants in chief (chief vassals) of the emperor and representatives of the imperial towns, was intended to cover the whole of Germany but was as powerless as its central authority. In the kingdoms of Spain, signs of development appeared very early, representatives of the towns being summoned in 1188 in Leon; by the mid-13th century, the presence of townsmen in representative assemblies was customary throughout the country. In Aragon magnates and knights were separate elements, but elsewhere the more usual pattern of the three estates of nobles, churchmen, and burgesses prevailed. The English gatherings to which the term parliament came increasingly to be applied differed in important respects. From early in the 13th century, the king had summoned his tenants in chief to meet at the same time the central courts of justice and finance were in session; thus, royal officers, innumerable representatives of the shires and boroughs, and the magnates were gathering at one time. By the end of the 13th century, it was becoming common for special representatives to be summoned from the shires and the boroughs to attend these parliaments; by the mid-14th century, their attendance had become the rule, but the clergy had largely withdrawn, except for the great ecclesiastical magnates who sat with the temporal lords, while the knights of the shire and the burgesses of the towns met together as a single body. Hence, there emerged a quite exceptional body with two chambers and a permanent representation of the non-noble landowners to contrast with the much more widespread model of the three estates.

English Parliament

The business undertaken by these gatherings was extremely various. For reasons essentially of convenience, it was customary to publish legislation at such assemblies, as church councils had done for centuries. In Aragon and Catalonia, Scandinavia and much of Germany, the consent of the assembly was required to legitimize such enactments; custom produced much the same effect in England, and the Aragonese kings of Sicily (who replaced the Angevins there during the War of the Sicilian Vespers, 1282–1302) proclaimed the same principle. Yet more frequently they were assembled to assent to taxation, sometimes, though never wholly, saving the prince the difficulty of negotiating with each town or community for a contribution to the common need; they continued to constitute also the most weighty public court for the hearing of great causes. Though summoned essentially to consent to decisions made by the king and his inner council, such assemblies necessarily possessed a potential power to refuse or demand a precedent redress of grievances; and as the powers of central government became greater, efforts to impose restraints on the prince in the name of the community became more widespread. Such restraints had long been recognized in theory, but the rise of these larger assemblies provided a focus for such claims that was more effective than the earlier courts of feudal tenants. Precocious efforts to subordinate royal government to the scrutiny of the magnates in Parliament, and even to require general assent to the appointment of officers of state, occurred in England between 1258 and 1265 and continued with some formal success in the 14th and 15th centuries. It was a successful struggle in the 14th century to require parliamentary consent to all extraordinary taxation that ensured the regular summons of such gatherings and made possible their use as an occasional forum of political discontent or even revolution, as in the reign of Richard II.

The States General in France made similar efforts at political control of the kingship in the crises associated with the capture of King John, between 1356 and 1358 and again in 1413. The Cortes of Spain secured an even larger measure of success; in 1287 the magnates extracted from Alfonso III of Aragon the Privilegio de la Unión, which conferred upon the Cortes even the power to depose an unjust king; here, too, assent to taxation was a prerogative of the assembly, though grants were made only by the burgesses. The weakness of these estates were very like those of the 15th-century councils of the church in their conflicts with the pope. Essentially they were occasions, rather than permanent bodies, which were either summoned by the king or—if gathered by any other means—were unrepresentative. Their control over the monarchy lasted only so long as they were in session, and exhaustion or particularism frustrated their long continuance. The kings could normally exploit either the localism or the diverging class interests of their estates to prevent any continuous supervision. In Catalonia in the 15th century, the Cortes possessed a Diputaçio del General, a standing committee to watch over the government, which not only granted but collected and spent any extraordinary taxation; but the forms of taxation divided the burgesses from the knights and nobles. In England there was less division of class interest but no effective or continuous supervision. In all cases, the resistance of the magnates was a precondition of large claims on behalf of the estates; yet a prolonged and successful magnate resistance destroyed that fusion of local and class interests upon which the estates depended for their vitality.

Cultural life. The communities of the early Middle Ages lived on intimate terms with a largely hostile environment. Storms on sea and land, floods, pestilences, and famine were constant hazards. Even in the more densely settled parts of Europe, impenetrable forest and trackless fen covered large areas. These areas were often believed to be the haunts of demons and, in fact, provided refuge for brigands and outlaws as well as wolves and wild beasts. The uncultivated moorlands and mountain passes were safely crossed only in haste and with company.

Pagan religious survivals

The religious beliefs of the pagan tribes were appropriate to this environment of apparently random disaster. The gods of the Germanic tribes were wanton and crafty, and fertility cults were probably widespread though intensely local. Formal belief in Odin, Thor, or Freyja gave way rapidly before the missionary fervour of Arian and Roman missionaries, but many pagan elements passed into

the superficial Christianity of the first converts. Devotion to local saints was often (even deliberately) based upon earlier pagan cults, while churches were built on the sites of temples or sacred groves. Early miracle stories often attribute markedly pagan qualities of willful jealousy or capricious good will to the saints. The mere possession of such wonder-working relics as the Holy Lance of the empire was supposed to confer a title to the crown and a formidable military authority to its holder. There was a brisk traffic in such objects among princes, while the sacred groves and wells of the earlier religion rapidly took on the name of some local saint for the peasant population. Augury, the sacrifice of cattle, and a host of other pagan rites continued throughout the period, though the ecclesiastical police system was more and more successful in sharpening the distinction between acceptable devotion and witchcraft.

Christian life at the parish level necessarily reflected such conditions. In a literate society it was the ceremonial performance of the sacraments that was of paramount importance; indeed, preaching was a late and occasional addition to the duties of the parish clergy. The blessing of harvests and houses, the averting of plague, fire, and invasion by regular formal intercession were the essential functions of the parish clergy. Correspondingly, communal disaster and outbreaks of anti-clericalism were closely connected. For the salvation of the individual the church required annual confession and very occasional taking of the sacrament. Private confession and private penance became the rule by 1100, but penance might be done by extensive travelling. By the 12th century, pilgrimages were made to the great international shrines of Jerusalem, Rome, Santiago de Compostela, or even St. Thomas of Canterbury, as well as to a host of lesser local shrines, insuring against the future or expiating past crimes. The Crusades, in part intended to restore the pilgrim road to Jerusalem, were themselves a means of securing full remission of sins for the participants. Their frequent failure or diversion to trivial or political ends made them a dying force after 1300, but the need for such concrete means of salvation continued (see CRUSADES). Papal jubilees and the organization of a system of indulgences secured by generous alms met this need, though offering many opportunities for abuse.

Until the 12th century, when Manichaeism spread through Italy and southern France, the orthodox church in the West was notably free of the doctrinal divisions that rent the Eastern Church. Eradicated by prolonged and bloody war, this heretical view had little later influence, but the source of its success—the wealth and pastoral inertia of the secular church—was to produce a multitude of dissident movements that laid stress either upon the defects of the propertied clergy or upon the direct illumination of the individual, notably by the reading of the Scriptures; already in the 12th century the Waldenses had adopted such views. In the late 14th century, John Wycliffe in England preached individual salvation and criticized the whole fabric of the church, sacramental and hierarchic. Wycliffe's doctrines met with little response in England but provided the starting point for the Hussite movement in Bohemia, where economic, political, and doctrinal revolution united to threaten the whole social and ecclesiastical order of eastern Europe.

In part this multiplication of dissent drew on the anticlericalism of the more educated laity and on wider knowledge of the Scriptures in vernacular versions, but it also coincided with the end of the period of monastic reform and development; while the fervour of St. Francis produced a new order to perform a vital function within the church, the devotional tendencies of the 14th and 15th centuries associated with the names of Meister Eckehart or Thomas à Kempis were largely devoid of institutional consequences. At the opposite extreme of this interior devotion was the proliferation of extravagant sects such as the recurring appearance of the flagellants in the years after the Black Death and its successors ravaged Europe. The prophetic form of much of the Scriptures, the frequency of disasters that appeared, at least locally, to portend the collapse of human society, and the general belief that the centuries after Christ's coming represented the last age of mankind produced a steady trickle of sects believing in the imminence of judgment or the new Jerusalem, encouraged by an easy distortion of the view of such academic prophets as Joachim of Fiore.

Communal organization

Against the hostility of their world, seen or unseen, medieval societies also fell back upon a wide variety of communities vowed to mutual support. The minimal privacy of the characteristic medieval household gave all houses a communal quality. It was common for several generations to live, eat, and sleep under one roof, which often enough covered the livestock too. The earliest forms of Germanic settlement and organization rested upon the kin group, the early codes all supposing that the kin had absolute responsibility for its members. In the unstable society of the early Middle Ages, the bonds of commendation by which men bound themselves to a protector provided a substitute for the security and social cohesiveness of the earlier kin group. The medieval impulse toward the formation of communities bound together by oaths may be seen also in the trading guilds of the communes, in the councils of kings and princes, and even in the societies of rebels or robber bands.

The tightly knit character of most medieval communities was a result of the relative immobility of the population. While the church and warfare were international occupations, in which a man might serve from one end of Europe to the other, and though merchants travelled with their wares across a multitude of frontiers, the bulk of the agricultural population rarely journeyed more than a few miles from their village, and many townsmen were equally confined. News travelled slowly and inaccurately, borne by pilgrims, peddlers, bailiffs, and beggars; marriage outside the village was unusual; and local dialects sharply contrasted. On the edge of this society, however, some adventurous travellers were covering huge distances. The Vikings sailed south to the Mediterranean, east and south to the Black Sea, and west to Iceland, Greenland, and the coast of North America. In the 13th century, the Venetian traveller Marco Polo crossed the breadth of Asia to China, where several Franciscan missionaries were to follow. In the 15th century, Portuguese seamen groped their way south along the African coast in pursuit of gold and a sea route to the Indies, rounding the Cape of Good Hope in 1488. In 1492 a Genoese seaman, in the service of Ferdinand and Isabella of Spain, crossed the Atlantic and returned; remarkable though Columbus' voyage was, the rapidity with which its significance was assessed and exploited provided proof of how much the European view of the world was changing.

Technological advances

Primitive though the vessels and instruments of these navigators were, they showed a marked advance over the equipment of their predecessors, an advance that was widespread and accelerating in the 14th and 15th centuries. From the 9th century on, water power was extensively harnessed to mill flour, replacing the infinitely laborious querns of past millennia; it was also at work driving the hammers of the ironworkers and the first fulling mills, which revolutionized the social patterns of cloth working. In the late 12th century, windmills also began to be used widely, while at about the same time changes in the forms of harness allowed horses to take over some of the ploughing and carting formerly performed by oxen. In the 13th century, mining techniques for the first time allowed the driving of deep but drained shafts beneath the surface to the richer iron, copper, tin, and lead of Bohemia, Sweden, or Cornwall. By 1500 the demand for charcoal for smelting and timber for shipbuilding was already pressing hard upon the once inexhaustible forest of an older Europe.

As a consequence of these changes, the use of iron became common, even in the implements and houses of the poor. Metal cooking pots, glass bottles, and glazed wheel-turned pottery came into widespread use. Houses of stone or even brick often now had the chimneys and glass windows formerly found only in palaces. Domestic furniture of these centuries survives to show that it was no longer confined to crude benches, tables, and chests but was more carefully fitted and elaborately decorated. In the 15th century the former cloth-working town of

Arras expanded its manufacture of the tapestry hangings for which it would be long famous, while silk, linen, and cotton came into wider use. Sumptuary laws showed how much wider a range of society was able to dress in the materials and fashions that had once been the hallmark of the highest rank.

Architecture

The earliest surviving medieval architecture is ecclesiastical; outside the area of continuing Byzantine occupation, these are small churches sometimes built out of the fragments of former Roman buildings and always derivative in style. Some of the greater Carolingian churches, however, were more ambitious in scale and conception, and the surviving 11th- and 12th-century Romanesque churches show the emergence of a wholly distinctive architectural convention in enormous and complex buildings. Gothic emerged in early 12th-century France and became the dominant style of the Latin West. In Italy the success of this style came late and was short-lived, for it was overtaken by the revived classical models before it had long taken root; but over much of northern Europe the Gothic survived into the 20th century.

The earliest surviving medieval domestic architecture is found in a few stone houses of the towns and in the modifications of fortresses to serve more domestic purposes. By the 13th century the stone palaces of princes were constructed with the skills first practiced in churches. Such planned towns as Aigues-Mortes foreshadowed the elaborate fortifications of many towns of the 15th century, within which numerous multistoried houses of wealthier merchants survive to the present. The single chamber of the lord's hall or the peasant's hut was increasingly giving way to the house divided into rooms, though in much of rural Europe there was little to distinguish the cabins of the poor of 1500 from those of a millennium earlier.

Recreation

The recreations of most ranks of society are poorly attested before the later Middle Ages. Among great men and warriors, the life of the hall and forest was pre-eminent. In the hall the long eating and heavy drinking of the lord's followers were favourite motifs of the heroic poetry once sung by household or itinerant poets, and feasts were an essential focus of social life. Coronations, weddings, and funerals were all celebrated with banquets. Pagan customs of punctuating the year with such events were either contested or taken over by the church; the gathering of the harvest and the celebration of Christmas were regular occasions of the kind, and from time to time the enthronement of a bishop produced a banquet of extraordinary splendour. The earliest texts stress the quantity (rather than the quality) of food and drink, but the more detailed later medieval records show that scarce or exotic foods and an extreme elaboration in their preparation were becoming the marks of ceremonial extravagance, the highest achievement of the pastrycook being such "subtleties" as the Holy Trinity surrounded by choirs of angels. Eating, drinking, and singing were supplemented by various forms of gambling, to which the Germanic peoples were so addicted that men were known to stake their own liberty on the fall of the dice and so condemn their descendants to slavery. Backgammon, chess, and (late in the period) card games were played by those with leisure and means.

The principal recreations of the nobility in the countryside were hunting and fighting, both of which were originally conducted in an extremely dangerous way. The main beasts chased were the deer, boar, wolf, and bear, with rabbits, hares, and foxes serving as a lesser challenge; all were usually pursued on horses with packs of hounds, the variety and qualities of which were discussed in numerous later medieval treatises. The support of the chase was an important economic and legal institution. The rearing of the lord's dogs and other services for his hunting were conditions of some tenures, and the preservation of hunting rights was a source of revenue and a cause of oppression. Another honourable sport was falconry. There was an appropriate bird for each rank of society, and in this sport women might participate more often than in the hunt.

The holding of sham battles for enjoyment and exercise became general in the 12th century, when the ransom of prisoners was an accepted means of support for skilled but landless knights. By the end of the century, the practice was so widespread as to attract the condemnation of the church and to represent, it was supposed, a threat to the stability of the state, since gatherings of armed men could readily become rebellions. In later centuries, this generalized melee gave way increasingly to highly formalized jousting between individual knights wearing armour especially adapted for a variety of possible rules of combat. In the 14th century this was still a possible means of capture, ransom, or death; by the end of the 15th it was almost entirely a sport.

For the less wealthy, the available entertainments were rarer and less elaborate but not perhaps much safer. Football, wrestling, or fighting with staves, regulated by few conventions, produced a heavy casualty rate, as did brawling in innumerable unregulated alehouses. Brewing was a cottage industry frequently engaged in by women, who sold their product subject to a seignorial right to examine the quality of the drink. In a society in which sugar was expensive even for kings and in which honey was prized but scarce, the thick beer and mead of the north was not only a focus of recreation but also an essential element of diet, as was the wine of the south.

Medieval music

Among the earliest of recorded arts was that of song, and a substantial body of medieval music has survived. The best known is the ecclesiastical plainsong, whose origins are attributed to the liturgical reforms of Pope Gregory I, which was elaborated into the complex polyphony of the 14th and 15th centuries, when for the first time the names of individual composers are recorded. The songs of troubadours and minnesingers of the 11th and 12th century show the rise of a sophisticated secular music that was increasingly accompanied by a variety of musical instruments. Well before 1500 the playing of music as an autonomous and often purely secular art was already established.

The decorative arts of painting and sculpture had undergone a similar progress, but here the revived interest in the secular products of classical antiquity and extensive lay patronage may have hastened the emergence of such arts from a wholly ecclesiastical setting to produce the flourishing schools of Italian and Flemish painting and the extensive sculpture of 15th-century Italy. In both cases the remarkable triumphs of the earlier period, the mosaics of Ravenna or Norman Sicily, for instance, or the sequences of sculpture across the west face of Autun or Laon were subordinate to a larger ecclesiastical unity—the liturgical drama of the church; the sculpture of Donatello and the painting of Botticelli were not, whatever the piety of their creators.

The rise of printing

The rise of the professional painter of pictures to be considered as a creator in his own right is largely matched by the decline of the illumination of manuscripts, which reached a peak of elaboration in the mid-15th century just as the earliest printed books were heralding the end of the role of the hand-copied manuscript. Printing on wooden blocks spread rapidly in the second half of the 15th century, and by 1500 presses were at work in every major country in Europe. Though the earliest printed books were extremely expensive and their purpose frequently liturgical, their potential importance was great because they could cater to the new market of the educated laity. These now included far more than the circle of great men who commissioned the splendid psalters and books of hours of the 14th and 15th centuries, which had overthrown a monastic monopoly of lavish manuscripts that had prevailed since the time of Cassiodorus and his copyists at Vivarium in the 6th century.

Education

The institutions and the subject matter of education had already proceeded according to a similar rhythm. The only schools of Europe in the 8th century, except perhaps in parts of northern Italy, were those attached often to monasteries and more rarely to bishoprics. Independent thought, even in theology, was extremely rare; theology remained under the long shadow of St. Augustine of Hippo, and the Greek learning and originality of John Scotus Erigena in the 9th century was little pursued. The rise of the new skills in dialectic in the 11th and 12th centuries produced two phenomena: first, a confidence in rational thought as a means of solving problems, especially those

raised by the conflict of authorities; and, second, a number of teachers whose exceptional talents attracted scholars from the farthest ends of Europe. The self-confidence and European reputation of Abelard reveal this movement at its most distinctive. Around such teachers grew up either religious communities such as that of Saint-Victor of Paris or the earliest universities. In the 12th century the lawyers of Bologna, the doctors of Salerno, and, above all, the theologians of Paris were becoming organized bodies governed by a chancellor; by the 13th century the universities possessed their own statutes regulating the arduous courses of study toward recognized degrees. The crown of studies was the pursuit of the highest knowledge, theology. The forms of 13th-century university study gave rise to the characteristic theological achievements of the period, the *summae* of the Dominican St. Thomas Aquinas and the Franciscan St. Bonaventure. The founding of universities received a new impetus at the end of the 14th and 15th centuries, when they spread into Scandinavia, Scotland, and eastern Europe. It was in these years that most secured a large independence from external ecclesiastical government, and in the councils of Constance and Basel the universities claimed a position of the highest authority.

Much of the earlier confidence in the capacity of intellectual endeavour according to the established forms of enquiry drained away in the 15th century. Logicians in the tradition of Duns Scotus and William of Ockham asserted the essential disparity of faith and reason; the canon law proved incapable of resolving the most pressing problems of ecclesiastical authority or of securing effective reforms; and the only literary forms that offered novelty and room for growth were the vernacular literature of the court and the classical poetry of the Humanists. Whatever changes occurred in education were generated not in the universities but in the schools of such Italians as Vittorino da Feltre near Mantua or of the Brethren of the Common Life in the Netherlands. These first insisted upon the effects of education, on the whole personality, where the numerous grammar, guild, and charitable schools had provided only a grounding in the mechanics of literacy. The retreat of the monopoly of the church in education stands beside the work of the explorers and the rise of the absolute monarchies as an important mark of the ending of medieval society.

(Ma.Br.)

The emergence of modern Europe, 1492–1648

THE CHRONOLOGY OF RENAISSANCE EUROPE

In the 15th century changes in the structure of European polity, accompanied by a new intellectual temper, suggested to such men as the philosopher and clerical statesman Nicholas of Cusa that the "Middle Age" had attained its conclusion and a new era had begun. The papacy, the symbol of the spiritual unity of Christendom, lost much of its prestige in the schism and the conciliar movement and became infected with the lay ideals prevailing in the Italian peninsula. In the 16th century the Protestant Reformation reacted against the worldliness and corruption of the Holy See, and the Catholic Church responded in its turn by a revival of piety known as the Counter-Reformation. While the forces that were to erupt in the Protestant movement were gathering strength, the narrow horizons of the Old World were widened by the expansion of Europe to America and the East.

In western Europe nation-states emerged under the aegis of strong monarchical governments, breaking down local immunities and destroying the unity of the European *Respublica Christiana.* Centralized bureaucracy came to replace medieval government. Underlying economic changes affected social stability. Secular values prevailed in politics, and the concept of a balance of power came to dominate international relations. Diplomacy and warfare were conducted by new methods. Permanent embassies were accredited between sovereigns, and on the battlefield standing armies of professional and mercenary soldiers took the place of the feudal array that had reflected the social structure of the past. At the same time scientific discoveries cast doubt on the traditional cosmology. The systems of Aristotle and Ptolemy, which had long been sanctified by clerical approval, were undermined by Copernicus, Mercator, Galileo, and Kepler.

The movement termed the Renaissance (see below) substituted new standards of culture and methods of thought for the Gothic styles and scholastic exercises of the past. The harmony of classical models was preferred to the disordered grandeur of medieval architecture, while in art pagan realism and the worship of the human form replaced religious symbolism, even in the representation of Biblical subjects. The revival of interest in ancient Greece and Rome was stimulated by an influx of Byzantine scholars, even before the conquest of the Constantinople by the Ottomans in 1453. The subsequent invention of the printing press popularized the rebirth of classical ideals and heightened the effect of scholarly criticism of religious and pseudohistorical dogmas. In both its rational, humanist aspect and its mystic, neoplatonic guise, the Renaissance expanded from its centre in the petty despotisms of Italy to all the confines of Europe.

Discovery of the New World. In the Iberian Peninsula the impetus of the counteroffensive against the Moors carried the Portuguese to probe the West African coastline and the Spanish to attempt the expulsion of Islām from the western Mediterranean. In the last years of the 15th century Portuguese navigators established the sea route to India, and within a decade had secured control of the trade routes in the Indian ocean and its approaches. Mercantile interests, crusading and missionary zeal, and scientific curiosity were intermingled as the motives for this epic achievement. Similar hopes inspired Spanish exploitation of the discovery by Christopher Columbus of the Caribbean outposts of the American continent in 1492. The Treaties of Tordesillas and Saragossa in 1494 and 1529 defined the limits of westward Spanish exploration and the eastern ventures of Portugal. The two states acting as the vanguard of the expansion of Europe had thus divided the newly discovered sea lanes of the world between them.

By the time of the Treaty of Saragossa, when Portugal secured the exclusion of Spain from the East Indies, Spain had begun the conquest of Central and South America. In 1519, the year in which Ferdinand Magellan embarked on the westward circumnavigation of the globe, Hernán Cortés launched his expedition against Mexico. The seizure of Peru by Francisco Pizarro and the enforcement of Portuguese claims to Brazil completed the major steps in the Iberian occupation of the continent. By the middle of the century the age of the conquistadores was replaced by an era of colonization, based both on the procurement of precious metal by Indian labour and on pastoral and plantation economies using imported African slaves. The influx of bullion into Europe became significant in the late 1520s, and from about 1550 it began to produce a profound effect upon the economy of the Old World.

Nation-states and dynastic rivalries. The organization of expansion overseas reflected in economic terms the political nationalism of the European states. This political development took place through processes of internal unification and the abolition of local privileges by the centralizing force of dynastic monarchies. In Spain the union of Aragon, Valencia, and Catalonia under John II of Aragon was extended to association with Castile through the marriage of his son Ferdinand with the Castilian heiress Isabella. The alliance grew toward union after the accession of the two sovereigns to their thrones in 1479 and 1474 respectively and with joint action against the Moors of Granada, the French in Italy, and the independent kingdom of Navarre. Yet at the same time provincial institutions long survived the dynastic union, and the representative assembly (Cortes) of Aragon continued to cling to its privileges when its Castilian counterpart had ceased to play any effective part. Castilian interest in the New World and Aragonese ties in Italy, moreover, resulted in the ambivalent nature of Spanish 16th-century policy, with its uneasy alternation between the Mediterranean and the Atlantic. The monarchy increased the central power by the absorption of military orders and the adaptation of the Hermandad, or police organization, and the Inquisition

for political purposes. During the reign of Charles I (the emperor Charles V) centralization was quickened by the importation of Burgundian conciliar methods of government, and in the reign of his son Philip II Spain was in practice an autocracy.

Other European monarchies imitated the system devised by Roman-law jurists and administrators in the Burgundian dominions along the eastern borders of France. In England and France the Hundred Years' War had reduced the strength of the aristocracies, the principal opponents of monarchical authority. The pursuit of strong, efficient government by the Tudors in England, following the example of their Yorkist predecessors, found a parallel in France under Louis XI and Francis I. In both countries revision of the administrative and judicial system proceeded through conciliar institutions, although in neither case did it result in the unification of different systems of law. A rising class of professional administrators came to fulfill the role of the king's executive. The creation of a central treasury under Francis I brought an order into French finances already achieved in England through Henry VII's adaptation of the machinery of the royal household. Henry VIII's minister, Thomas Cromwell, introduced an aspect of modernity into English fiscal administration by the creation of courts of revenue on bureaucratic lines. In both countries the monarchy extended its influence over the government of the church. The unrestricted ability to make law was established by the English crown in partnership with Parliament; in France the representative States General lost its authority and sovereignty reposed in the king in council. Supreme courts (*parlements*) possessing the right to register royal edicts, imposed a slight and ineffective limitation on the absolutism of the Valois kings. The most able exponent of the reform of the judicial machinery of the French monarch was Charles IX's chancellor, Michel de L'Hôpital, but his reforms in the 1560s were frustrated by the anarchy of the religious wars. In France the middle class aspired to ennoblement in the royal administration and mortgaged their future to the monarchy by investment in office and the royal finances. In England, on the other hand, a greater flexibility in social relations was preserved, and the middle class engaged in bolder commercial and industrial ventures.

Territorial unity under the French crown was attained through the recovery of feudal appanages (alienated to cadet branches of the royal dynasty) and, as in Spain, through marriage alliances. Brittany was regained in this way, although the first of the three Valois marriages with Breton heiresses also set in train the dynastic rivalry of Valois and Habsburg. When Charles VIII of France married Anne of Brittany he stole the bride of the Austrian archduke and future emperor Maximilian I and also broke his own engagement to Margaret of Austria, Maximilian's daughter by Mary of Burgundy. Margaret's brother Philip, however, married Joan, heiress of Castile and Aragon, so that their son eventually inherited not only Habsburg Germany and the Burgundian Netherlands but also Spain, Spanish Italy, and America. The dominions of Charles V thus encircled France and incorporated the wealth of Spain overseas. Even after the division of this vast inheritance between his son, Philip II of Spain, and his brother, the emperor Ferdinand I, the conflict between the Habsburgs and the French crown dominated the diplomacy of Europe for over a century.

The principal dynastic conflict of the age was less unequal than it seemed, for the greater resources of Charles V were offset by their cumbrous disunity and by local independence. In the Low Countries he was able to complete the Seventeen Provinces by new acquisitions, but, although the coordinating machinery of the Burgundian dukes remained in formal existence, Charles's regents were obliged to respect local privileges and to act through constitutional forms. In Germany, where his grandfather Maximilian I had unsuccessfully tried to reform the constitution of the Holy Roman Empire, Charles V could do little to overcome the independence of the lay and ecclesiastical princes, the imperial knights, and the free cities. The revolts of the knights (1522) and the peasantry (1525), together with the political disaggregation imposed by the Reformation, rendered the empire a source of weakness. Even in Spain, where the rebellion of the *comuneros* took place in 1520–21, his authority was sometimes flouted. His allies, England and the papacy, at times supported France to procure their own profit. France, for its part, possessed the advantages of internal lines of communication and a relatively compact territory, while its alliance with the Turk maintained pressure on the Habsburg defenses in southeast Europe and the Mediterranean. Francis I, however, like his predecessors Charles VIII and Louis XII, made the strategic error of wasting his strength in Italy, where the major campaigns were fought in the first half of the century. Only under Henry II was it appreciated that the most suitable area for French expansion lay toward the Rhine.

Turkey and eastern Europe. A contemporary who rivalled the power and prestige of Francis I and Charles V was the ruler of the Ottoman Empire, the sultan Süleyman I the Magnificent (1520–66). With their infantry *corps d'élite* (the janissaries), their artillery, and their cavalry, or *sipahis,* the Ottomans were the foremost military power in Europe, and it was fortunate for their Christian adversaries that Eastern preoccupations prevented them from taking full advantage of Western disunity. A counterpoise was provided by the rise of the powerful military order of the Ṣafavids in Persia—hostile to the orthodox Ottomans through their acceptance of the heretical Islāmic cult of the Shī'ah. Ottoman strength was further dissipated by the need to enforce the allegiance of Turkmen begs in Anatolia and of the chieftains of the Caucasus and Kurdistan and to maintain the conquest of the sultanate of Syria and Egypt by Süleyman's predecessor, Selim I. Süleyman himself overran Iraq and even challenged Portuguese dominion of the Indian Ocean from his bases in Suez and Basra. The Crimean Tatars acknowledged his suzerainty, as did the corsair powers of Algiers, Tunis, and Tripoli. His armies conquered Hungary in 1526 and threatened Vienna in 1529. With the expansion of his authority along the North African coast and the Adriatic littoral, it seemed for a time as if the Mediterranean, like the Black Sea and the Aegean, might become an Ottoman lake.

Though it observed the forms of an Islāmic legal code, Turkish rule was an unlimited despotism, suffering from none of the financial and constitutional weaknesses of Western states. With its disciplined standing army and its tributary populations, the Ottoman Empire feared no internal threat except the periods of disputed succession, which continued to occur despite a law empowering the reigning sultan to put to death collateral heirs. It was not unusual for the sultan to content himself with the overlordship of frontier provinces. Moldavia and Walachia were for a time held in this fashion, and in Transylvania the *voivode* John Zápolya gladly accepted Süleyman as his master in return for support against Ferdinand of Austria.

Despite the expeditions of Charles V against Algiers and Tunis, and the inspired resistance of Venice and Genoa in the war of 1537–40, the Ottomans retained the intiative in the Mediterranean until several years after the death of Süleyman. The Knights of St. John were driven from Rhodes and Tripoli and barely succeeded in retaining Malta. Even after the combined fleet of Spain, the papacy, Venice, and Genoa had crushed the Turkish armament in 1571 in the Battle of Lepanto, the Ottomans took Cyprus and recovered Tunis from the garrison installed by the allied commander, Don John of Austria. North Africa remained an outpost of Islām and its corsairs continued to harry Christian shipping, but the Ottoman Empire did not again threaten Europe by land and sea until late in the 17th century.

In eastern Europe the states of Poland, Lithuania, Bohemia, and Hungary were all loosely associated at the close of the 15th century under rulers of the Jagiellon dynasty. In 1569, three years before the death of the last Jagiellon king of Lithuania–Poland, these two countries merged their separate institutions by the Union of Lublin. Thereafter the Polish nobility and the Roman Catholic faith dominated the Orthodox lands of Lithuania and held the frontiers against Muscovy, the Cossacks, and the Tatars. Bohemia and the vestiges of independent Hungary

were regained by the Habsburgs as a result of dynastic marriages, which the emperor Maximilian I planned as successfully in the East as he did in the West. When Louis II of Hungary died fighting the Ottomans at Mohács in 1526, the Archduke Ferdinand of Austria obtained both crowns and endeavoured to affirm the hereditary authority of his dynasty against aristocratic insistence on the principle of election. In 1619 Habsburg claims in Bohemia became the ostensible cause of the Thirty Years' War, when the Diet of Prague momentarily succeeded in deposing Ferdinand II.

In the 16th century eastern Europe displayed the opposite tendency to the advance of princely absolutism in the West. West of the Carpathians, and in the lands drained by the Vistula and the Dnestr, the landowning class achieved a political independence that weakened the power of monarchy. The towns entered a period of decline, and the propertied class, though divided by rivalry between the magnates and the lesser gentry, everywhere reduced their peasantry to servitude. In Poland and Bohemia the peasants were reduced to serfdom in 1493 and 1497 respectively, and in free Hungary the last peasant rights were suppressed after the rising of 1514. The gentry, or *szlachta,* controlled Polish policy in the Sejm (parliament), and when the first Vasa king, Sigismund III, tried to reassert the authority of the crown after his election in 1587, the opportunity had passed. Yet despite the anarchic quality of Polish politics, the aristocracy maintained and even extended the boundaries of the state. In 1525 they compelled the submission of the secularized Teutonic Order in East Prussia, resisted the pressure of Muscovy, and pressed to the southeast, where communications with the Black Sea had been closed by the Ottomans and their tributaries.

Farther to the east the grand duchy of Muscovy emerged as a new and powerful despotism. Muscovy, and not Poland, became the heir to Kiev during the reign of Ivan III the Great in the second half of the 15th century. By his marriage with the Byzantine princess Sofia (Zoë) Palaeologus, Ivan also laid claim to the traditions of Constantinople. His capture of Novgorod and repudiation of Tatar overlordship began a movement of Muscovite expansion, which was continued by the seizure of Smolensk by his son Vasily (Basil) III and by the campaigns of his grandson Ivan IV the Terrible (1533–84). The latter destroyed the khanates of Kazan and Astrakhan and reached the Baltic by his conquest of Livonia from Poland and the Knights of the Sword. He was the first to use the title of tsar, and his arbitrary exercise of power was more ruthless and less predictable than that of the Ottoman sultan. After his death Muscovy was engulfed in the Time of Troubles, when Polish, Swedish, and Cossack armies devastated the land. The accession of the Romanov dynasty in 1613 heralded a period of gradual recovery. Except for occasional embassies, the importation of a few Western artisans, and the reception of Tudor trading missions, Muscovy remained isolated from the West. Despite its relationship with Greek civilization, it knew nothing of the Renaissance. Though it experienced a schism within its own Orthodox faith, it was equally untouched by Reformation and Counter-Reformation, the political consequences of which convulsed western Europe in the second half of the 16th century.

Reformation and Counter-Reformation. In a sense the Reformation was a protest against the secular values of the Renaissance. No Italian despots better represented the profligacy, the materialism, and the intellectual hedonism that accompanied these values than did the three Renaissance popes, Alexander VI, Julius II, and Leo X. Among those precursors of the reformers who were conscious of the betrayal of Christian ideals were figures so diverse as the fiery Ferraran monk Savonarola, the Spanish statesman Cardinal Jiménez, and the Humanist scholar Erasmus.

The corruption of the religious orders and the cynical abuse of the fiscal machinery of the church provoked a movement that at first demanded reform from within and ultimately chose the path of separation. When the Augustinian monk Martin Luther protested against the sale of indulgences in 1517, he found himself obliged to extend his doctrinal arguments until his stand led him to deny the authority of the pope. In the past, as in the controversies between pope and emperor, such challenges had resulted in mere temporary disunity. In the age of nation-states the political implications of the dispute resulted in the irreparable fragmentation of clerical authority.

Luther had chosen to attack a lucrative source of papal revenue, and his intractable spirit obliged Leo X to excommunicate him. The problem became of as much concern to the emperor as it was to the pope, for Luther's eloquent writings evoked a wave of enthusiasm throughout Germany. The reformer was by instinct a social conservative and supported existing secular authority against the upthrust of the lower orders. Although the Diet of Worms accepted the excommunication in 1521, Luther found protection among the princes. In 1529 the rulers of electoral Saxony, Brandenburg, Hesses, Lüneberg, and Anhalt signed the "protest" against an attempt to enforce obedience. By this time Charles V had resolved to suppress Protestantism and to abandon conciliation. In 1527 his mutinous troops had sacked Rome and secured the person of Pope Clement VII, who had deserted the imperial cause in favour of Francis I after the latter's defeat at the Battle of Pavia. The sack of Rome proved a turning point both for the emperor and the Humanist movement that he had patronized. The Humanist scholars were dispersed, and the initiative for reform then lay in the hands of the more violent and uncompromising party. Charles V himself experienced a revulsion of conscience that placed him at the head of the Catholic reaction. The empire he ruled in name was now divided into hostile camps. The Catholic princes of Germany had discussed measures for joint action at Regensburg in 1524; in 1530 the Protestants formed a defensive league at Schmalkalden. Reconciliation was attempted in 1541 and 1548, but the German rift could no longer be healed.

Lutheranism laid its emphasis doctrinally on justification by faith and politically on the God-given powers of the secular ruler. Other Protestants reached different conclusions and diverged widely from one another in their interpretation of the sacraments. In Geneva, Calvinism enforced a stern moral code and preached the mystery of grace with predestinarian conviction. It proclaimed the separation of church and state, but in practice its organization tended to produce a type of theocracy. Huldreich Zwingli and Heinrich Bullinger in Zürich taught a theology not unlike Calvin's but preferred to see government in terms of the godly magistrate. On the left wing of these movements were the Anabaptists whose pacifism and mystic detachment were paradoxically associated with violent upheavals.

Lutheranism established itself in north Germany and Scandinavia and for a time exercised a wide influence both in eastern Europe and in the west. Where it was not officially adopted by the ruling prince, however, the more militant Calvinist faith tended to take its place. Calvinism spread northward from the upper Rhine and established itself firmly in Scotland and in southern and western France. Friction between Rome and nationalist tendencies within the Catholic Church facilitated the spread of Protestantism. In France the Gallican Church was traditionally nationalist and anti-papal in outlook, while in England the Reformation in its early stages took the form of the preservation of Catholic doctrine and the denial of papal jurisdiction. After periods of Calvinist and then of Catholic reaction, the Church of England achieved a measure of stability with the Elizabethan religious settlement.

In the years between the papal confirmation of the Jesuit order in 1540 and the formal dissolution of the Council of Trent in 1563, the Catholic Church responded to the Protestant challenge by purging itself of the abuses and ambiguities that had opened the way to revolt. Thus prepared, the Counter-Reformation embarked upon recovery of the schismatic branches of Western Christianity. Foremost in this crusade were the Jesuits, established as a well-educated and disciplined arm of the papacy by Ignatius Loyola. Their work was made easier by the Council of Trent, which did not, like earlier councils, result in the diminution of papal authority. The council condemned such abuses as pluralism, affirmed the traditional practice

in questions of clerical marriage and the use of the Bible, and clarified doctrine on controversial issues such as the nature of the Eucharist, divine grace, and justification by faith. The church in this way made it clear that it was not prepared to compromise; and with the aid of the Inquisition and the material resources of the Habsburgs, it set out to reestablish its universal authority. It was of vital importance to this task that the popes of the Counter-Reformation were men of sincere conviction and initiative, who skilfully employed diplomacy, persuasion, and force in the offensive against heresy. In Italy, Spain, Bavaria, Austria, Bohemia, Poland, and the southern Netherlands (the future Belgium), Protestant influence was destroyed. As the movement gathered strength in the age of religious warfare, its course became confused with the national ambitions of Philip II of Spain.

The Wars of Religion. Germany, France, and the Netherlands each achieved a settlement of the religious problem by means of war, and in each case the solution contained original aspects. In Germany the territorial formula of *cuius regio, eius religio* applied—that is to say, in each petty state the population had to conform to the religion of the ruler. In France, the Edict of Nantes in 1598 embraced the provisions of previous treaties and accorded the Protestant Huguenots toleration within the state, together with the political and military means of defending the privileges that they had exacted. The southern Netherlands remained Catholic and Spanish, but the Dutch provinces formed an independent Protestant federation in which republican and dynastic influences were nicely balanced. Nowhere was toleration accepted as a positive moral principle, and seldom was it granted except through political necessity.

There were occasions when the Wars of Religion assumed the guise of a supranational conflict between Reformation and Counter-Reformation. Spanish, Savoyard, and papal troops supported the Catholic cause in France against Huguenots aided by Protestant princes in England and Germany. In the Low Countries, English, French, and German armies intervened; and at sea Dutch, Huguenot, and English corsairs fought the Battle of the Atlantic against the Spanish champion of the Counter-Reformation. In 1588 the destruction of the Spanish Armada against England was intimately connected with the progress of the struggles in France and the Netherlands.

Behind this ideological grouping of the powers, national, dynastic, and mercenary interests generally prevailed. The Lutheran duke Maurice of Saxony assisted Charles V in the first Schmalkaldic War in 1547 in order to win the Saxon electoral dignity from his Protestant cousin, John Frederick; while the Catholic Henry II of France supported the Lutheran cause in the second Schmalkaldic War in 1552 to secure French bases in Alsace. John Casimir of the Palatinate, the Calvinist champion of Protestantism in France and the Low Countries, maintained an understanding with the neighbouring princes of Lorraine, who led the ultra-Catholic Holy League in France. In the French conflicts Lutheran German princes served against the Huguenots, and mercenary armies on either side often fought against the defenders of their own religion. On the one hand, deep divisions separated Calvinist from Lutheran; and, on the other hand, political considerations persuaded the moderate Catholic faction, the Politiques, to oppose the Holy League. The national and religious aspects of the foreign policy of Philip II of Spain were not always in accord. Mutual distrust existed between him and his French allies, the family of Guise, because of their ambitions for their niece Mary Stuart. His desire to perpetuate French weakness through civil war led him at one point to negotiate with the Huguenot leader, Henry of Navarre (afterward Henry IV of France). His policy of religious uniformity in the Netherlands alienated the most wealthy and prosperous part of his dominions. Finally, his ambition to make England and France the satellites of Spain weakened his ability to suppress Protestantism in both countries.

In 1562, seven years after the Peace of Augsburg had established a truce in Germany on the basis of territorialism, France became the centre of religious wars which endured, with brief intermissions, for 36 years. The political interests of the aristocracy and the vacillating policy of balance pursued by Henry II's widow, Catherine de Médicis, prolonged these conflicts. After a period of warfare and massacre, in which the atrocities of St. Bartholomew's Day (1572) were symptomatic of the fanaticism of the age, Huguenot resistance to the crown was replaced by Catholic opposition to the monarchy's policy of conciliation to Protestants at home and anti-Spanish alliances abroad. The revolt of the Holy League against the prospect of a Protestant king in the person of Henry of Navarre released new forces among the Catholic lower classes, which the aristocratic leadership was unable to control. Eventually Henry won his way to the throne after the extinction of the Valois line, overcame separatist tendencies in the provinces, and secured peace by accepting Catholicism. The policy of the Bourbon dynasty resumed the tradition of Francis I, and under the later guidance of Cardinal Richelieu the potential authority of the monarchy was realized.

In the Netherlands the wise Burgundian policies of Charles V were largely abandoned by Philip II and his lieutenants. Taxation, the Inquisition, and the suppression of privilieges for a time provoked the combined resistance of Catholic and Protestant. The House of Orange, represented by William I the Silent and Louis of Nassau, acted as the focus of the revolt; and in the undogmatic and flexible personality of William, the rebels found leadership in many ways similar to that of Henry of Navarre. The sack of the city of Antwerp by mutinous Spanish soldiery in 1576 (three years after the dismissal of Philip II's autocratic and capable governor, the Duque de Alba) completed the commercial decline of Spain's greatest economic asset. In 1579 Alessandro Farnese, duke of Parma, succeeded in recovering the allegiance of the Catholic provinces, while the Protestant north declared its independence. French and English intervention failed to secure the defeat of Spain, but the dispersal of the Armada and the diversion of Parma's resources to aid the Holy League in France enabled the United Provinces of the Netherlands to survive. A 12-year truce was negotiated in 1609, and when the campaign began again it merged into the general conflict of the Thirty Years' War. It was during this last phase of the struggle that Amsterdam came to replace Antwerp as the trading centre of Europe.

(J.H.McM.S.)

Thirty Year's War. *Western Europe.* The later wars of aggression, waged by Louis XIV of France and by Napoleon I, tend to obscure the fact that from the middle of the 16th to the middle of the 17th century France was the victim of a deliberate policy of encirclement by the Habsburg powers. In the south, possession of Roussillon gave Spain a firm foothold north of the Pyrenees. In the southeast, the Republic of Genoa was a Spanish satellite and the Duchy of Milan was Spanish territory. Genoa and Milan guaranteed a short and safe line of communication from Spain to Austria proper as well as to the Spanish possessions along the eastern borders of France, namely, Franche-Comté (contiguous with Austrian Alsace) and the Low Countries (Luxembourg Hainaut, and Artois on the French frontier, with Flanders, Brabant, and the rest of the Netherlands in their immediate hinterland). After the successes of the Dutch rebellion, from 1572, the overland route from Genoa to Brussels was the more important because the final link of the chain around France had snapped after the death of Mary I of England (1558): the fate of the Armada in 1588 had demonstrated the insecurity of the northern sea route from Spain to the Netherlands.

Habsburg encirclement of France

The breaking of this Habsburg stranglehold was the foremost task of French statesmanship. It was first undertaken when Henry IV of France resisted the attempt by the Habsburgs, in 1609, to acquire Jülich-Cleves-Mark-Berg, as this attempt threatened to deal the deathblow to the independence of the United Provinces of the Netherlands and to lay open France's northern frontier. Henry's policy, interrupted by his assassination (1610) and by the pro-Spanish appeasement of the French regency under Marie de Médicis, was resumed by Cardinal de Richelieu,

in power from 1624 until his death in 1642, and was continued by his successor, Cardinal Mazarin. France, at first too weak to wage open war against the Habsburgs, began by making treaties with powers hostile to them (or to their satellites in Germany or in eastern Europe), namely, with the United Provinces (1624), with Sweden (1631), and with Russia (1632); but the defeat of the United Provinces' ally Denmark (Peace of Lübeck, 1629) the withdrawal of Russia (Peace of Polyanov, 1634), the defeat of the Swedes (Battle of Nördlingen, 1634), and the defection of Sweden's German allies (Peace of Prague, 1635) forced Richelieu to declare open war on Spain (1635), after which Catalonia and Portugal (1640) and later Great Britain (1657) also joined the French camp. The modest French military successes were surpassed at the conference table: the Peace of Westphalia (1648) established the preponderance of France and France's allies throughout central and northern Europe; the formation of the League of the Rhine (1658) gave France a decisive voice in the affairs of the empire; and the Peace of the Pyrenees (1659) enlarged and secured France's frontiers and marked the end of Spain as a great power.

Twelve years' Truce of Antwerp

For the Dutch, the period of the Thirty Years' War forms part of their Eighty Years' War against Spain (1568–1648). The 12 years' Truce of Antwerp (April 9, 1609) can now be seen as the de facto recognition of the independence of the United Provinces; but contemporary Dutch statesmen foresaw that the Spanish crown would not easily abandon the hope of reducing the rebellious heretics and recovering its richest European province. In fact, Philip IV of Spain, whose accession in 1621 coincided with the expiry of the truce, immediately renewed the war. But whereas at the outbreak of the rebellion Spain had been the leading maritime, commercial, and colonial power, superiority in all these spheres had passed to the Netherlands. The only weakness of the maritime republic was its lack of land forces; but the States-General had no difficulty in enlisting foreign mercenaries or paying subsidies to foreign princes. Whereas the interest of Spain demanded the localization of every conflict so that resources might be concentrated against the Dutch, the interest of the Dutch was served best by extending the war so as to engage Spain's power away from their borders. Religious and constitutional affinities gave the Calvinist Dutch republicans additional zest in their support (1) of the Calvinist elector John Sigismund of Brandenburg in his struggle for Jülich; (2) of the semirepublican nobles of Bohemia in their fight against Austrian absolutism; and (3) of the generals who, from 1621, continued the Bohemian War, nominally on behalf of the Calvinist elector Frederick V of the Palatinate. But practical political considerations made the United Provinces the focal point of every anti-Habsburg coalition: the Franco-Dutch Treaty of Compiègne (June 20, 1624) was the prelude of the Treaty of The Hague (1625), which effectively brought Denmark into the war against the emperor Ferdinand II; and, after Denmark's defeat, Dutch diplomacy and subsidies aided Richelieu in enlisting Gustavus II Adolphus of Sweden as the military champion of the anti-Habsburg cause (1630–32).

From the middle of the 1630s, the French superseded the Dutch as leaders of the anti-Habsburg struggle. The conquest of Alsace and of Breisach by France's protégé Bernhard of Saxe-Weimar in 1638 and the secession of Portugal in 1640 shook Spain into preferring to come to an arrangement with the Dutch; and the latter, having smashed the Spanish Navy in the English Channel (1639) and off Pernambuco (1640), thereby putting an end to Spanish overseas expansion, now considered France and the Portuguese colonial empire greater rivals to their mercantile interests than Spain. Disregarding the alliance with France, which they had made in 1635, the Dutch on January 30, 1648, signed a separate peace with Spain.

Struggle for Baltic supremacy

Northern Europe. Viewed from northern Europe, the first half of the 17th century comprises the attempts of the two Scandinavian kingdoms, Denmark and Sweden, to obtain what the diplomatic language of the time called the *dominium maris Baltici* ("lordship of the Baltic Sea"); *i.e.*, the possession of the leading Baltic ports through whose customs sheds the raw materials of the North German, Polish, and Russian hinterlands found their way into the West.

Denmark's repeated attempts to overpower Sweden failed, and the Danish outposts in the eastern Baltic were lost. King Christian IV's attempt to make himself supreme in Lower Saxony by acquiring the German bishoprics of Bremen, Verden, Minden, and Hildesheim led to the Danish War (1625–29), in which the superior generalship of his adversaries outweighed the subsidies granted to him by the Dutch and the British. The Peace of Lübeck (1629) finished Denmark as a European power of consequence.

Denmark's decline was emphasized by the simultaneous rise of Sweden. The Peace of Knäred (1613) made only transitory concessions to Christian IV of Denmark, who had attacked Sweden in 1611. Thenceforward, King Gustavus II Adolphus systematically began to close the ring of Swedish possessions around the Baltic. After preliminary gains at the expense of Muscovite Russia, whose access to the Baltic he closed (1617), he successfully attacked the Baltic provinces of Poland and of Poland's vassal, Prussia (1621–29). Besides pursuing this Baltic ambition, he may also have been aspiring to supplant his Catholic cousin, Sigismund III Vasa, on the Polish throne. Any such design, however, had to be abandoned: on the one hand the Swedish nobility objected to it; on the other, the king's Baltic policy was challenged by the advance of imperial armies along the southern shore from the west, which drew him into intervening in German affairs. The Swedish-Polish Truce of Altmark (1629) was followed by alliance with France (Treaty of Bärwalde, January 23, 1631).

The alliance with France survived the death of Gustavus Adolphus and remained the sheet anchor of his successors' position until Sweden's collapse as a great power (1721). Sweden reaped the fruits of this policy; first in the Peace of Brömsebro at the end of the Swedish-Danish War of 1643–45; then in the Peace of Westphalia (1648); and finally in the Peace of Roskilde (1658), with Denmark in the course of the First Northern War. By the end of this period Sweden had achieved dominance in the Baltic area.

Eastern Europe. The first half of the 17th century also embraces the first attempt by Poland to force Orthodox Muscovy (Russia) into the orbit of Latin Christianity as well as the first attempt on the part of the House of Romanov to enter the comity of western Europe. Poland, tied to the Habsburgs by religion and tradition, hostile to Sweden by religion and dynastic rivalry, and implacably hostile to Muscovy by history and religion, exploited the "troubles" that convulsed Russia after the death of Boris Godunov (1605). The experiment of ruling Russia through a Polish puppet, the first False Dmitry, failed in 1606 when the usurper succumbed in an outbreak of Russian nationalism and Orthodox fanaticism. But Poland came nearest to making Russia a satellite when in 1610 Sigismund III had his son, the future Wladyslaw IV of Poland, elected as tsar: the Polish dictatorship in Moscow lasted two years.

Russia's "Time of Troubles"

In 1609, the first Russo-Swedish alliance had been concluded by the tsar Vasily III Shuysky; and, despite Sweden's later aggressions, the new Romanov dynasty (from 1613) continued to regard Sweden as Russia's natural ally against Poland. England and the United Provinces acted as intermediaries at the Russo-Swedish Peace of Stolbova (1617), and 13 years later French envoys brought about Muscovy's indirect support of Gustavus Adolphus' war in Germany. Russia, however, did not want to be involved directly in Germany, and the Russo-Polish War of Smolensk (1632–34) remained a sideshow. After Gustavus Adolphus died, Swedish-Russian relations cooled, as the Swedish chancellor Axel Oxenstierna correctly assessed Russia as potentially more dangerous than Poland.

The Russo-Polish Treaty of Polyanov (1634), which included Wladyslaw IV's final resignation of his claim to the Russian throne, freed Poland to resume hostilities against Sweden and, by thus tying down Swedish troops, contributed to the Swedish disaster at Nördlingen.

Germany. All these European conflicts affected Germany more often and more deeply than any of the other contestants. Germany was the only country where the Reformation had resulted in a permanent split into three

Dynastic rivalry

religious factions—Catholic, Lutheran, and Calvinist. As these religious divisions largely marched with political frontiers, they were sustained and aggravated by dynastic rivalries, such as that between the Catholic Wittelsbachs in Munich (Bavaria) and the Calvinist Wittelsbachs in Heidelberg (the Palatinate) or that between the Calvinist Kassel branch and the Lutheran Darmstadt branch of the House of Hesse. In turn, these political and religious dissensions were partly exacerbated and partly overlaid by constitutional problems: the Holy Roman emperor, a Habsburg and a Catholic, wanted to establish monarchical absolutism in the empire; the electors, Catholics and Protestants alike, wanted to maintain what they called the "electoral preeminence" in the empire's affairs; and lastly, the other princes wanted to overthrow the ascendancy of both the emperor and the electors, so as to obtain complete freedom of action for themselves.

Each of the rivals found it easy and profitable to call in some foreign power. On the whole, the emperor relied on his Spanish cousins; the Protestant towns and smaller princes relied on Sweden, on France, and on the Dutch; Catholic Bavaria, with its satellites Cologne, Liège, Münster, Paderborn, and Hildesheim, usually blackmailed the emperor, while inclining to France and opposed to Spain; Lutheran Saxony disliked all foreign entanglements, but regarded the Calvinists as worse than the Catholics and generally found it most profitable to side with the emperor (since repeated endeavours to rally a neutralist "third force" never succeeded); and Protestant Brandenburg was weak and irresolute until 1640, when Frederick William, "the Great Elector," began to play off emperor, Swedes, Dutch, Poles, and French with such skill that by 1660 he had raised Brandenburg-Prussia to the rank of a great German and minor European power.

Great Britain. From the death of Queen Elizabeth I (1603) to the passing of the first Navigation Act by Cromwell's Parliament (1651), Great Britain's influence upon European affairs was negligible. The European policy of the government was vacillating between Spain, France, and the United Provinces, while Parliament and public opinion were enthusiastically anti-Spanish but unwilling to transform brave words into hard cash.

James I originally wanted to continue Elizabeth's pro-French and pro-Dutch policy, but assented to peace with Spain (1604). In 1613 he gave his daughter Elizabeth in marriage to Frederick V of the Palatinate, leader of the Protestant anti-Habsburg faction in Germany; but twice, in 1617 and in 1623, he tried to obtain a Catholic Spanish Habsburg princess for his son Charles before betrothing him, in 1624, to a Catholic French Bourbon.

James had at least not dissuaded his son-in-law from accepting the Bohemian crown (1619), but his dislike of the Czech nobility's republican sympathies and religious radicalism, combined with his desire to appease the Spaniards, prevented him from giving tangible support to Frederick during the war in Bohemia. It was only when Frederick had lost the Palatinate and when the Spanish marriage project had collapsed that a cautious parliamentary grant for the preparation of war against Spain gave expression to the English public's lively sympathy for the Protestant cause. An expedition to the Palatinate was hopelessly bungled.

Charles I, on succeeding his father, undertook the war against Spain and also associated himself with the Dutch in the Treaty of The Hague (December 9, 1625) for the support of Christian of Denmark against the emperor and the Catholic League in Germany. But the naval expedition against Cádiz (1625) was a failure; and Charles failed to pay Denmark the promised £360,000 per annum, after the first installment of £46,000. Soon afterward, he squandered ships, troops, and money on the disastrous expedition of 1627 to help the Huguenot rebels of La Rochelle against his brother-in-law, Louis XIII of France. Thereafter Charles's devious foreign policy was hampered by lack of funds, which forced him to conclude peace both with France (1629) and with Spain (1630), and by lack of interest in foreign affairs on the part of his most capable adviser, Strafford. In 1636 and again in 1641, Charles offered to the Habsburgs an alliance against the Dutch in return for the restoration of the Palatinate to his nephew Charles Louis; but the Habsburgs declined this offer because Bavaria objected to it. Richelieu likewise in the same years preferred for France an understanding with Bavaria to a possible rapprochement with England.

Religious aspects. Religious issues were insolubly interwoven with political and constitutional problems. For Germany the whole question stemmed from the Peace of Augsburg of 1555, a compromise between the Catholic and Lutheran Estates (members) of the empire. Neither party was satisfied, neither meant to abide by it. It fixed the religious frontiers as they had existed in 1552; *i.e.*, at the lowest ebb of Protestant fortunes. It gave the sovereign lay princes the right to change with their own religion that of their subjects; but it prohibited any Catholic bishopric or free city from adhering in future to the Lutheran creed. It expressly excluded Calvinists from the toleration granted to Lutherans. Above all, it failed to lay down any criteria by which the many doubtful points might be interpreted or to create any machinery to enforce its observance.

Peace of Augsburg

One of the many points left in suspense by the Peace of Augsburg concerned the right of the Protestant gentry in town and country under Catholic sovereigns to keep Protestant pastors and hold Protestant services. This issue played a part in the revolts of the Austrian and Bohemian nobility against their Habsburg overlords, which began in 1609 and ended with the Bohemian War in 1620. The ecclesiastical reservation that forbade the "reform" of Catholic bishoprics prevented Cologne from turning Protestant together with its archbishop, Gebhard, in 1583. Gebhard's deposition was the basis of the greatness of the House of Wittelsbach: Bavarian princes held the archbishopric of Cologne continuously from 1583 to 1761, the bishoprics of Hildesheim and of Liège almost uninterruptedly throughout the same period, and the bishoprics of Paderborn and of Münster with but few more interruptions. Thus the undisputed leadership of Catholic Germany by the Wittelsbach Maximilian I of Bavaria was grounded on the pre-eminence of his house not only in the south but also in the northwest of Germany, where Jülich and Berg in the hands of his brother-in-law fortified his house.

The distribution of the religious parties in the empire during the first half of the 17th century, according, be it understood, to the sovereign's choice of religion, was approximately as follows: the whole of northern, northeastern, and central Germany, with the exception of the bishopric of Hildesheim and the abbey of Fulda, was Protestant; northwestern, western, and southeastern Germany, with the exception of the Rhenish Palatinate in the west, was Catholic, as were all the Habsburg dominions (Austria and Tirol, Bohemia-Moravia and Silesia, a portion of Swabia). Württemberg in Swabia, Ansbach and Bayreuth in Franconia, and nearly all the free cities, even those in otherwise Catholic districts, adhered to the Protestant faith. The nobility and townspeople in Bavaria and the Habsburg countries, nearly all of whom had been Protestants, had been recatholicized or expelled by about 1625.

The historian must not doubt the strength and sincerity with which the champions of the contesting factions and their humblest followers believed in the exclusive truth of their church: the emperor Ferdinand II, Maximilian of Bavaria, Tilly, and Richelieu were devout Catholics; Gustavus Adolphus, Bernhard of Weimar, the electors of Brandenburg, Saxony, and the Palatinate were unswerving Protestants; Wallenstein is perhaps the only personality of note who seems to have been entirely indifferent to religious distinctions, astrology having largely supplanted Christianity as his creed. Motives other than religious zeal, however, were the factors determining political allegiance.

The subordination of religion to politics

Pope Urban VIII and the cardinals Richelieu and Mazarin were unbending in their opposition to Catholic Spain and to the Catholic emperor. The Lutheran elector of Saxony and the Lutheran landgrave of Hesse-Darmstadt were firm adherents of the emperor so long as circumstances permitted. Lutheran Saxony and Catholic Bavaria were consistently anti-Spanish and anti-Swedish. Calvinist Hesse-Kassel was Lutheran Sweden's most reliable ally and, together with Bavaria, mostly pro-French. The struggle for

the Baltic brought the Lutheran Gustavus Adolphus into conflict with Orthodox Russia, with Catholic Poland, and with Lutheran Denmark; but his fight in Germany made him the ally of Catholic France and of Orthodox Russia. Maximilian of Bavaria extended his power at the expense alike of the Lutheran cities in Swabia and in Franconia, of the Calvinist elector Palatine, and of the Habsburg emperor and did so mostly in conjunction with Lutheran Saxony and Catholic France. The bishops of Würzburg and of Bamberg were the first to abandon the Catholic interest at the peace congress of Münster. Protestant as well as Catholic princes, prelates, and cities were susceptible to French subsidies. Of none of the belligerents can it be said that religious motives were responsible for any major decision. The secret debates (1629–30) in the Swedish council about entry into the war are revealing: the security and defense of Sweden and the conquest of Germany were declared the war aims; the king was expressly warned against speaking of a war of religion since France might take umbrage.

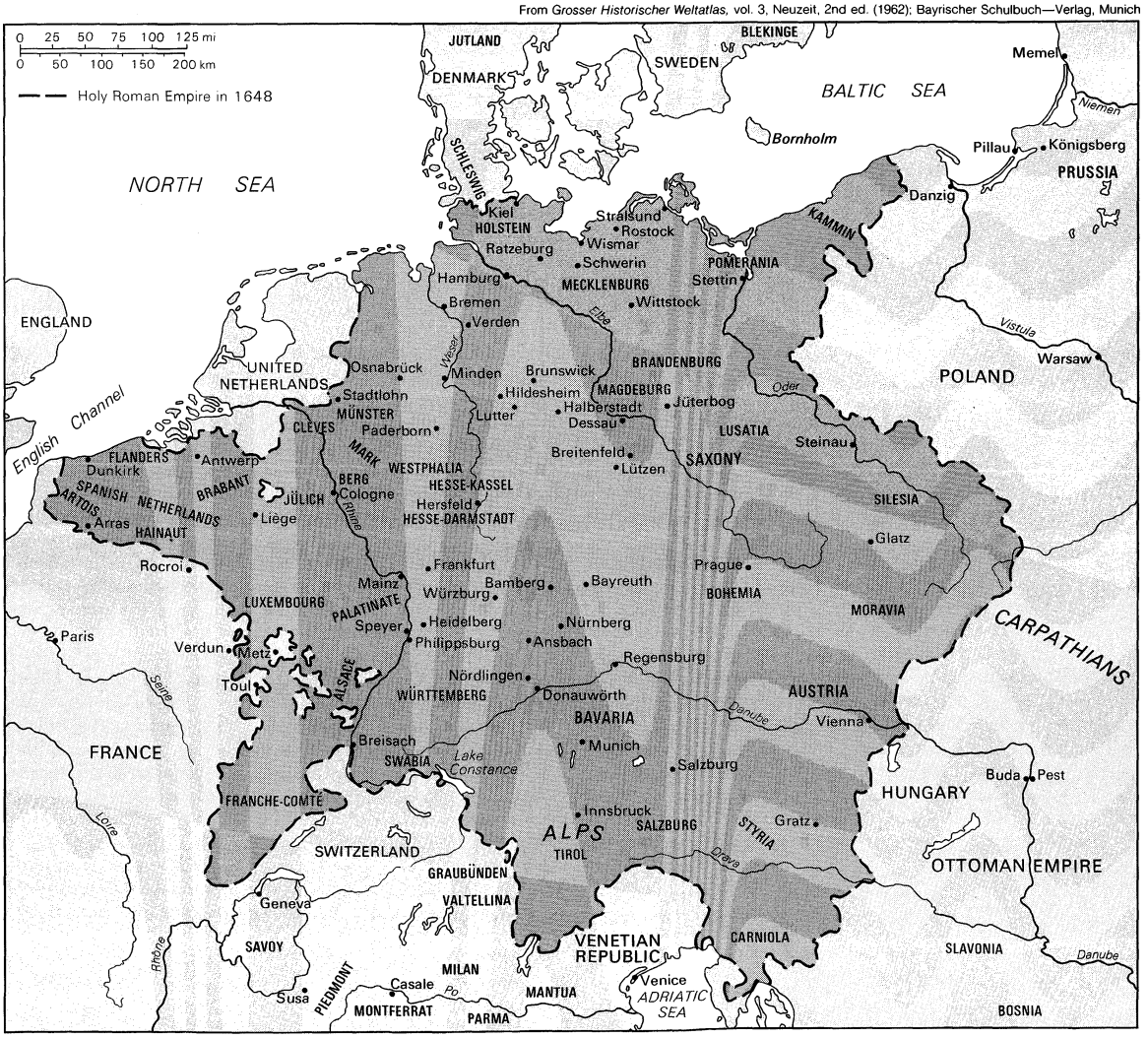
Success of monarchy

The suppression of Protestantism by Ferdinand II in Austria and in Bohemia and by Richelieu in France was undertaken mainly on political considerations. The Protestants not only infringed the spiritual unit of the country (a tenet still held by every government) but also were the chief opponents of royal absolutism and, not without cause, were suspected of republican tendencies on Dutch and Swiss models. Their defeat therefore paved the way for that spiritual and administrative uniformity that became the hallmark of 17th- and 18th-century monarchy.

Similarly, the contest of the emperor against the Protestant princes and cities was to a large extent a contest to decide whether the empire was to be a monarchy, with the emperor at its head and the princes as his vassals, or a federation of more or less independent princes, with the emperor as its titular president. So long as Ferdinand II was willing to concede to the six electoral princes their traditional prerogatives in the government of the empire, these electors (including Lutheran Saxony and Calvinist Brandenburg) sided with him against the lesser princes. When Ferdinand in 1629 unmasked his real intentions by the Edict of Restitution, the electors, led by Catholic Bavaria, turned against him. The Peace of Westphalia brought about the triumph of the lesser princes, with whom the electors now identified themselves against any recurrence of imperial centralism.

In fact, the Peace of Westphalia proved the interdependence of the problems, religious and constitutional, political and economic, that again and again were used as interchangeable counters. The recognition of the adherents of the Reformed (Calvinist) religion as being of equal right with those of the Lutheran faith was largely a byproduct of the dispute over Pomerania between Brandenburg and Sweden. The emperor sacrificed the north German bishoprics to the Protestant claimants (Sweden, Brandenburg, Saxony, Mecklenburg, and Brunswick) and Alsace and the bishoprics of Metz, Toul, and Verdun to France; and in return Sweden and France renounced their claims to Austrian territory in Silesia and in Swabia respectively and assented to the exclusion of the Austrian Protestants from the treaty's clauses on restitution and toleration. The attitude of the Catholic princes to Pope Innocent X's protest against any formal agreement with the Protestants is characteristic of the trend toward excluding denominational considerations from political decisions: the first secret intimation that the nuncio gave to the emperor and to the

From *Grosser Historischer Weltatlas*, vol. 3, Neuzeit, 2nd ed. (1962); Bayrischer Schulbuch—Verlag, Munich



Sites associated with the Thirty Years' War.

Catholic princes in November 1647 was answered with evasive excuses; and the formal protocol handed over on Christmas Eve did not even receive an answer from any prince, temporal or ecclesiastical. A special "anti-protest clause" in the peace treaty was signed by all parties: it declared Innocent's condemnation of the peace invalid and ineffective and thus demonstrated the complete emancipation of secular politics from ecclesiastical tutelage. The religious questions left unanswered by the Peace of Augsburg in 1555 were solved in 1648 in the modern spirit of the "reason of state."

Catholic League and Protestant Union

War of the Jülich Succession (1609–14). The settlement that the Peace of Augsburg had formulated for Germany was shaken in 1607–08 when Maximilian I of Bavaria annexed and recatholicized the Lutheran city of Donauwörth. Some of the German Protestant states concluded a military alliance, the Union, on May 14, 1608. The death of Duke John William of Jülich on March 25, 1609, opened a crucial question of succession over which war broke out. On July 10, 1609, a Catholic military alliance, the League, was formed.

The wealth of John William's lands—Cleves, Jülich, Mark, and Berg—and their strategic position attracted not only the powerless legitimate heirs, namely, John Sigismund of Brandenburg and Wolfgang Wilhelm of Palatinate-Neuburg but also every powerful neighbour. The Habsburgs wished to install an Austrian prince who would have placed the lands at the disposal of the Spaniards in their struggle with the Dutch and would incidentally have counteracted the predominance of Bavaria in northwestern Germany. France, the United Provinces, and England naturally objected to any strengthening of the Spaniards in that region. The assassination of Henry IV of France (May 14, 1610) prevented the war from becoming a European contest, though Spanish, Austrian, and Dutch troops repeatedly invaded the duchy in pretended support of one or the other of the heirs. The Protestant Union, which John Sigismund had joined by turning Calvinist, and the Catholic League, of which Wolfgang Wilhelm had become a member by entering the Roman Church and marrying a sister of Maximilian, brought about a compromise that excluded all foreign claims (October 24, 1610) and led to the partition of the inheritance between Brandenburg and Neuburg (Treaty of Xanten, 1614).

Bohemian and Palatine War (1618–23). This war began with the insurrection, in 1618, of the Bohemian and Austrian Estates against the future emperor Ferdinand II, whose intention was to impose absolutist rule and to enforce the Catholic Counter-Reformation. The Bohemian nobles toyed with the idea of setting up a republic in order to secure the support of the Dutch but eventually, in August 1619, elected as their king the elector Frederick V of the Palatinate in the hope of obtaining the aid of the Protestant Union and Great Britain. Their expectations proved fallacious. Frederick failed to discipline the haughty Bohemian nobles or to rally the Bohemian townsmen and peasants (whom he did nothing to relieve from harsh oppression by their feudal lords) and was abandoned by his German allies as well as by his father-in-law, James I of Great Britain. Ferdinand, although in a very weak position, succeeded in buying the support of Maximilian of Bavaria and the Catholic League at a heavy price in money and land. In a campaign of a few months, the League's troops under Johann Tserclaes von Tilly crushed the Austrian Estates and broke the rule of Frederick in the

Battle of the White Mountain

Battle of the White Mountain, near Prague (November 8, 1620). The conquest of Glatz (Klodzko) on October 25, 1622, completed the subjugation of the lands of the Bohemian crown.

In the meantime, the Spaniards from their bastions in Luxembourg and in Franche-Comté had since 1620 been overrunning the Rhenish Palatinate, where Tilly joined them in 1662, after he and Maximilian had in 1621 subdued the Upper Palatinate (north of Bavaria and west of Bohemia). Tilly also overcame the badly coordinated actions of several German princelings and generals who, in the pay of the United Provinces, Denmark, and England, still upheld Frederick's cause, notably, Ernst von Mansfeld and Christian of Brunswick. Whereas the Protestant Union had dissolved itself (May 1621), Lutheran Saxony, which had never adhered to the Union, was won over to the emperor by the offer of Lusatia. The defeat of Christian of Brunswick by Tilly at Stadtlohn (August 6, 1623) left the situation as follows: the emperor was in undisputed control of his Austrian and Bohemian territories; Maximilian of Bavaria, who was created elector, had become the leading power in southern and northwestern Germany; the Spaniards were in possession of the Rhenish Palatinate; and Frederick was a landless exile, living on the bounty of the Dutch.

The Valtellina

Struggle for Graubünden (1620–39). The Valtellina, leading from the northern frontier of the Duchy of Milan through the Alps toward Tirol, opened the shortest and safest land route between Spanish Italy and the Austrian territories in Germany, whence the Spaniards could reinforce their troops in the Netherlands and in the Palatinate. It belonged to Graubünden, or the Grisons, a union of leagues in loose relations with the Swiss. Spain and Austria obviously wanted to bring the Union's lands under their own control; France and Venice naturally opposed the forging of this link, which would have completed encirclement of both of them by the Habsburg powers. The political and military issues were here poisoned by religious, personal, local, and clannish rivalries, of which Georg Jenatsch was the stormy centre. A Spanish occupation of the Valtellina (1620), reinforced by an Austrian incursion into the other territories of Graubünden (1621), provoked French and Venetian protests, which led to the nominal "deposit" of the Valtellina in the hands of the papacy (1623)—in fact a veiled means of prolonging Spanish control there. Then Richelieu ventured on the first French occupation of Graubünden (1624–26); but the internal weakness of France, where Richelieu's regime was challenged on the one hand by the pro-Spanish faction of Catholic zealots (the Parti Dévot) and on the other by the Huguenots, made it impossible to maintain this indirect attack on Spain. The Treaty of Monzón (1626) made Graubünden into a sort of Franco-Spanish protectorate, with papal troops in occupation. During the War of the Mantuan Succession, the Habsburgs again overran the country (1629–31). A final French occupation (from 1635) was ended in 1639 by a reaction of Graubünden against the French. Eventually the Peace of Milan (September 3, 1639), between Spain and Graubünden, brought the country into virtually complete dependence on Spain.

Swedish–Polish War (1621–29). From 1611, Gustavus Adolphus had taken advantage of the "Time of Troubles" and its aftermath in Muscovite Russia; and by the Peace of Stolbova (1617) he had acquired Karelia and Ingria, which together constituted the land bridge between Swedish Finland and Swedish Estonia, so that the Gulf of Finland was converted into a Swedish lake. He then turned against Poland (1621). His conquest of Livonia gave him Riga, the chief Baltic port (though the Swedish nobles possessed themselves of the largest land properties in the country, to the detriment of the Swedish crown); and from the Duchy of Prussia, Poland's vassal, he took the ports of Memel and Pillau, the latter commanding the approach to Königsberg. Sweden's position in the Baltic, however, was eventually endangered by the course of the Danish War, as the imperial general Wallenstein conquered Mecklenburg (with the ports of Wismar and Rostock) and threatened moreover to conquer Pomerania; and in summer 1628 Swedish forces were sent to help the port of Stralsund, which Wallenstein was besieging. Finally, the Truce of Altmark (September 25, 1629) was concluded between Sweden and Poland, largely thanks to French mediation. Livonia and the Prussian ports were left in Sweden's possession.

Protestant coalition

Danish War (1625–29) and the Edict of Restitution. Christian IV of Denmark had observed strict neutrality during the Bohemian War despite repeated Dutch attempts to enlist his help on behalf of Frederick of the Palatinate. In 1624, however, he began to contemplate offensive action. His motives were rivalry of Sweden in the Baltic regions and the wish for Danish supremacy in the estuaries of the Elbe and Weser. The king's election, in spring 1625, as director of the Lower Saxon Circle of the empire

furnished him with the legal pretext for interference in Germany. The Treaty of The Hague (December 9, 1625), in the negotiation of which the British statesman George Villiers, 1st duke of Buckingham, played a major part, promised British and Dutch subsidies for a military effort by Denmark on Frederick's behalf, in cooperation with the German generals who had already tried to uphold the latter's cause; and this Protestant coalition could expect the sympathy of the Habsburgs' other enemies—prince Gábor Bethlen of Transylvania, the Ottoman Turks, and also Catholic France.

The coalition's plan seems to have envisaged a fourfold advance: Christian of Brunswick was to overpower the Wittelsbach bishoprics and duchies in Westphalia and in the lower Rhineland; Christian IV of Denmark was to make himself master of Lower Saxony; Ernst von Mansfeld, generalissimo of the coalition, was to press forward into Bohemia, Silesia, and Moravia; and Bethlen was to sally forth from Hungary and join forces with Mansfeld.

On the opposite side were ranged the veteran troops of the Catholic League under Tilly, who was to deal with Christian of Brunswick and with Christian IV of Denmark, and the imperial army of Wallenstein, who was to repel Mansfeld and Bethlen.

On both sides the generals were on bad terms with one another and never effectively coordinated their efforts. But Tilly and Wallenstein had the advantage of fighting on interior lines and could therefore tackle their adversaries in turn. The coalition, on the other hand, was weakened by the faithless policy of Charles I of Great Britain, who failed to honour his financial obligations and let Buckingham embark on a foolish expedition against France in aid of the Huguenots of La Rochelle (June–November 1627). Wallenstein forced Mansfeld out of northern Germany (Battle of Dessau, April 25, 1626); and Mansfeld, though he held a strong position in Silesia, neither made full use of anti-Habsburg movements (apart from discontent in the Bohemian lands, a peasants' revolt broke out in Austria in May) nor effected a junction with Bethlen and the latter's Turkish auxiliaries. Wallenstein outmanoeuvred Bethlen, and Mansfeld died on his way to Venice (November 29, 1626), whence he hoped to receive fresh subsidies. Isolated Danish troops maintained themselves in Silesia until the autumn of 1627, when Wallenstein overwhelmed them.

Meanwhile, Tilly was favoured by the untimely death of Christian of Brunswick (June 16, 1626) and by the military incompetence of Christian IV of Denmark. The Danish Army suffered a crushing defeat in the Battle of Lutter (August 27, 1626). In 1627 Tilly pursued the beaten Danes into Holstein; and Wallenstein, who joined him, continued the pursuit into Jutland. Mecklenburg, whose dukes had sided with Christian, was given by the emperor to Wallenstein, who immediately set about obtaining bases on the Baltic for an imperial navy. His attempt to add the Pomeranian port of Stralsund to his Mecklenburg ports of Wismar and Rostock led to the intervention of Gustavus Adolphus.

Wallenstein's contact with maritime affairs changed his whole outlook. His first sight of large river barges on the Oder had made him believe that they were oceangoing capital ships; but he now realized the importance of sea power and overseas trade and reassessed the political position that the emperor might take up with regard to the Hanse towns, Denmark, the Netherlands, England, and Spain. Denmark, which he wished to draw into the emperor's interest, profited by his mediation: at the Peace of Lübeck (May 22, 1629) Christian IV had only to renounce further participation in the affairs of the empire.

The defeat of the anti-Habsburg coalition raised the position of the emperor to its greatest height since Charles V's victory over the League of Schmalkalden in 1547.

Edict of Restitution

Ferdinand II's Edict of Restitution (March 6, 1629), prescribing the recovery by the Catholics of all ecclesiastical lands in which Protestantism had been established since 1552, was more than an act of reparation of the damage suffered by the Catholic Church since Luther's time: it was an unambiguous assertion of the imperial prerogative in all matters pertaining to the constitutional structure of the empire. At the Electoral Diet of Regensburg (1630) the opposition, led by Lutheran Saxony (anxious to retain its threatened acquisitions) and by Catholic Bavaria (determined to counteract the emperor's aggrandizement) forced Ferdinand to dismiss Wallenstein and so to make himself defenseless at the very moment when Gustavus Adolphus had landed on German soil.

War of the Mantuan Succession (1628–31). The death of Vincenzo II Gonzaga, duke of Mantua and Montferrat, on December 26, 1627, led to the first direct clash between the Habsburgs and France, albeit on a secondary theatre of war. The claims of the legitimate Gonzaga heir, Charles, duc de Nevers, whom France supported, were overridden by Ferdinand II, supreme lord of the imperial fiefs in Italy; and Spain lent military aid to Ferdinand so as to keep a French vassal away from the approaches to Milan. Richelieu, however, was soon stronger. The Huguenot rebels of La Rochelle capitulated to Louis XIII (autumn 1628); Savoyard forces, which attempted to block the French king's way into Piedmont, were defeated at Susa (March 6, 1629); and the Peace of Susa with England (April 14), followed by the submission of the Huguenot rebels in Languedoc to the Peace of Alais (June 28), secured the French rear. The Habsburg forces withdrew from Casale in October 1630; and Richelieu, who in November triumphed over the pro-Spanish faction in France (the so-called Day of Dupes), achieved full success. By the Treaty of Cherasco (1631), the French candidate was installed in Mantua. Meanwhile, Savoy had ceded the fortress of Pinerolo to France; and Pope Urban VIII, determined opponent of the Habsburgs, annexed Urbino, another vacant imperial fief. The Austro-Spanish monopoly in Italy was broken.

Swedish War (1630–35) and the Peace of Prague. The suspension of hostilities between Sweden and Poland freed Gustavus Adolphus to intervene in Germany. After landing at Usedom (July 6, 1630), he quickly occupied the whole duchy of Pomerania and restored Wallenstein's duchy of Mecklenburg to its hereditary dukes. He concluded the Treaty of Bärwalde (January 23, 1631) with France, whereby he was to receive annual subsidies of 1,000,000 livres to enable him to campaign for the restoration of the "liberty" of the German princes "oppressed" by the emperor. In these circumstances an immediate objective for Gustavus was to stop the enforcement of the Edict of Restitution on Magdeburg, which commanded a strategic crossing point on the Elbe; but his advance into central Germany was impeded by the hardly veiled hostility of the electors George William of Brandenburg and John George I of Saxony. Saxony tried at the Convent, or Conference, of Leipzig (February 1631) to establish a neutral party between the emperor and Sweden, which proved unacceptable to either. Though Gustavus stormed the Brandenburg fortress of Frankfurt an der Oder (April 13), thereby both securing his left flank against the Poles and intimidating the two electors, the delay led to the fall of Magdeburg to Tilly (May 20). Brandenburg and Saxony now yielded to the Swedish threats, especially as Tilly imprudently invaded Saxony and treated it as enemy country. Alliances with Brandenburg (June 20) and with Saxony (September 11) protected Gustavus' rear and virtually rendered the electorates Swedish satellites. The defeat of Tilly at Breitenfeld (September 17) and the conquest of Prague (November 15) by the Saxons under Hans Georg von Arnim opened southern Germany to the Swedes.

Battle of Breitenfeld

The smaller German Protestant princes, such as Bernhard of Saxe-Weimar, flocked to Gustavus' standards; by April 1632 Gustavus had advanced as far as Munich, and his armies had reached Lake Constance to the south and Mainz to the west. The archbishopric of Mainz and the Rhenish Palatinate were placed under Swedish administration; the bishoprics of Würzburg and Bamberg were given to Bernhard as a Swedish fief under the name of Duchy of Franconia. The king's war aims gradually widened: his agreements with the German princes became more onerous, and the stipulations that the Treaty of Bärwalde had made for the protection of France's friends and of the Catholic religion were treated lightly. Gustavus did not aim at the imperial crown, but his secret negotiations with Wallenstein show that he thought

of placing the Habsburg dominions under the rule of Swedish puppets.

The recall of Wallenstein to the leadership of the imperial army completely changed the situation: Gustavus was manoeuvred out of southern Germany and shortly afterward was killed in the Battle of Lützen (November 16, 1632). While the conduct of Swedish military operations then fell to Johan Banér and subsequently Lennart Torstenson, the direction of policy was undertaken by the chancellor, Axel Oxenstierna, who, by the Treaty of Heilbronn (April 23, 1633), consolidated the Swedish alliance with the German princes, albeit without Brandenburg and Saxony. Wallenstein's victory at Steinau in Silesia (October 11) and Bernhard's capture of Regensburg (November 14) cancelled each other out, and Wallenstein's machinations to make himself the arbiter of affairs caused the emperor to have him murdered (February 25, 1634); but the Swedish defeat at Nördlingen (September 5–6, 1634) led to the dissolution of the League of Heilbronn and to the open defection of most German princes, led by Saxony, from the Swedish cause.

The Peace of Prague (May 30, 1635) reconciled the emperor and nearly all the German opponents of the Edict of Restitution, which it modified by making the year 1627 the criterion of rightful possession of ecclesiastical lands (instead of 1552). Thenceforth, Sweden played only a subordinate part in the war. The leading role fell to France.

War of Smolenski (1632–34). From *c.* 1628, Gustavus Adolphus and Richelieu had been trying to bring about alliances with Russia, Turkey, Transylvania, the Crimean Tatars, and the Ukrainian Cossacks, who were to engage the emperor and Poland on their eastern frontiers. Concerted action, however, proved difficult to achieve. But Moscow gave Gustavus considerable indirect aid by selling to Sweden, on a cash-and-carry basis at an artificially low price, cereals that the Swedes resold in Amsterdam at considerable profit: between 1628 and 1633 the Swedes bought Russian grain at a cost of 100,000 talers per annum and sold it at 400,000 talers per annum (the latter sum being equal in amount to the direct subsidy received annually from France). Russo-Swedish military cooperation was eventually achieved, largely through the efforts of Sweden's envoy Alexander Leslie (later earl of Leven in the peerage of Scotland): the Russians invaded Poland in autumn 1632 and besieged Smolensk, while Gustavus Adolphus moved eastward and ordered his general Carl Gustav von Wrangel to prepare an offensive from his bases in Prussia.

Swedish-Russian cooperation

The great scheme came to nought. The deaths of Gustavus Adolphus and of the Moscow patriarch Philaret (October 1633), who had been the two chief architects of Swedish-Russian cooperation, removed the mutual goodwill as the basis for a real alliance. Turkey, moreover, was committed in a war with Persia; the Tatars of the Crimea turned against Muscovy instead of Poland; and the revolt of the Ukrainian Cossacks against their Polish overlords was delayed. Finally, an insurrection of the peasants in central Russia forced Tsar Michael's government to conclude the Peace of Polyanow with Poland (June 14, 1634).

French and Swedish War (1635–48). The near-collapse of the Swedish system in Germany in 1634–35 forced Richelieu to abandon his cautious policy of nonintervention. He concluded offensive and defensive alliances with the United Provinces (February 8, 1635) and with Sweden (Treaty of Compiègne, April 28), sent a French army to the Valtellina (March–April), and declared war on Spain (May 19). He then secured an alliance with Savoy and Parma (League of Rivoli, July 11); mediated the 20-year Truce of Stuhmsdorf between Sweden and Poland (September 12); and took the best of the German generals still serving Sweden, Bernhard of Saxe-Weimar, into French pay (October 27).

In 1636 the invasion of northern France from the Spanish Netherlands by Ottavio Piccolomini, whose capture of Corbie (August 15) threatened to expose Paris, was outweighed by a series of Swedish victories over Saxon and imperial forces, culminating in Banér's defeat of Melchior von Hatzfeldt at Wittstock (October 4), which reestablished Swedish supremacy in northern and central Germany. A careless northward movement of the imperial commander in chief, Matthias Gallas, in 1637, laid southern Germany open to the French; and Bernhard of Saxe-Weimar, having overrun Alsace, in 1638 conducted a brilliant campaign in the course of which he took the key fortresses of Rheinfelden (March 23), Freiburg (April 6), and Breisach (December 17). Before Bernhard's death, the incompetence of the French led to Piccolomini's great victory of Thionville (June 7, 1639), which, however, was to be the last success of the Austro-Spanish armies. The outbreak, in 1640, of revolution both in Catalonia and in Portugal compelled the Spaniards to limit their commitments outside the Iberian Peninsula. The capitulation of the great stronghold of Arras to the French (August 9, 1640) endangered the Spanish position in the Netherlands, and a Spanish counteroffensive into Champagne ended with the French victory at Rocroi (May 19, 1643), won by the young duc d'Enghien (Prince de Condé).

In Germany the various commanders—French, Swedish, Bavarian, and imperial—waged war almost on each one's own responsibility: no coherent pattern can be found in those campaigns, which were nearly always small-scale raids with limited objectives. The Swedish victory over Saxon and imperial forces at Breitenfeld (November 2, 1642) and the capture of the Little Town quarter of Prague by the Swedes (July 26, 1648) are—besides the Swedish campaign against Denmark—the most notable military events of these years. They were overshadowed by the diplomatic activity that began in 1640 and ended in 1648 with the Peace of Westphalia.

During the French and Swedish War, death had removed three great figures from the international scene: the emperor Ferdinand II on February 15, 1637, leaving the succession to his son Ferdinand III; Richelieu on December 4, 1642; and Louis XIII of France on May 14, 1643. Richelieu's place was taken by Cardinal Mazarin during the regency for Louis XIV.

Swedish–Danish War (1643–45). The precarious situation of the imperial cause after Breitenfeld and the Danish jealousy and fear of Sweden brought about an understanding between Ferdinand III and Christian IV; and Sweden's decision to wage a preventive war against Denmark gave the emperor a respite, since the Swedish army under Torstenson had to abandon its march on Vienna (September 1643) and to turn northward instead. In a lightning campaign (December 1643–January 1644) Torstenson conquered Schleswig-Holstein and Jutland. When an imperial army under Gallas came to the succour of the Danes, Torstenson at once marched against him, destroyed his army at Jüterbog (November 23, 1644), and invaded Bohemia, where another imperial army was wiped out at Jankov (March 6, 1645). By the Peace of Brömsebro, signed on August 23, 1645, Denmark ceded Jämtland and Härjedalen on the Norwegian frontier, Halland on the Kattegat, and the Baltic islands of Gotland and Ösel to Sweden.

Peace of Westphalia (1648). Treaties signed in the Westphalian towns of Münster and Osnabrück terminated both the Eighty Years' War between Spain and the United Provinces of the Netherlands (January 30, 1648) and the war between France, Sweden, and the German Protestants on the one side and the emperor and the German princes on the other (October 24). The Peace of Prague (1635) between the emperor Ferdinand II and the majority of the German princes had proved abortive: it expressed too much the temporary ascendancy of the emperor, took little account of Sweden, and completely disregarded France, which at that very moment openly took up arms against Spain and the emperor. No general peace was possible without the participation of Sweden, France, and Spain; and in 1640 the parties began in earnest to prepare the summoning of a peace congress. The emperor Ferdinand III entered into secret negotiations with Sweden in Hamburg; the Imperial Diet demanded a universal congress instead of bilateral transactions. The renewal of the Franco-Swedish alliance on June 30, 1641, included the stipulation that two congresses be held simultaneously in neighbouring Westphalian towns: the Catholic envoys were to meet in Münster, the Protestant envoys, in Osnabrück.

In 1643–44 Sweden and France sent out the first invitations, and the congress began to take shape. The slow and tortuous negotiations about procedure and substance, protocol and ceremonial, the admission or rejection of envoys and mediators—all these fumblings, which went on right to the actual signing of the treaties—were mostly due not to obstruction but to inexperience. By trial and error the congress had to explore and define the methods of modern international diplomacy. The 150 representatives (about 110 Germans and 40 foreigners) lacked any precedent for their tasks, starting, as it were, from scratch every time a fresh topic turned up. They gradually developed a kind of esprit de corps, which cut across political and religious frontiers and contributed not a little to the realistic settlement, which was satisfactory to all but a few diehards.

The negotiators

Though the plenipotentiaries were bound fairly narrowly by instructions from their governments, a few figures stand out to whom the successful outcome was largely due. The imperial ambassador, Maximilian Graf von Trauttmansdorff, outshone the rest; he had been the architect of the Peace of Prague and became the main author of the Peace of Westphalia. He and his chief secretary, Isaac Volmar, were converts to the Roman faith and therefore understood the opposing positions and did their best to check the firebrands in the Catholic camp. France was brillantly represented by the duc de Longueville (Henry d'Orléans) and his duchesse (Anne Geneviève de Bourbon-Condé), but as members of the princely faction they were suspect to Mazarin. Mazarin also mistrusted France's principal actual negotiator, Claude de Mesmes, comte d'Avaux, whereas the latter's equally skillful but more ruthless colleague, Abel Servien, enjoyed the cardinal's confidence; their mutual antagonism, however, in no way impaired their efficiency as representatives of the French crown. Similar dissensions rent the Swedish mission: Johan Oxenstierna, the chancellor's son, stood for the policy of conquest sponsored by the Swedish nobility, whereas Johan Adler Salvius, a gifted and experienced diplomat of humble birth, sided with the young queen Christina in wishing for peace at almost any price. The Spanish ambassadors, the conde de Peñaranda (Gaspar de Bracamonte) and Antonius Brun, succeeded in terminating the war with the Dutch (January 30, 1648) with great sacrifice to Spanish power but without loss to Spanish honour.

The Brunswick counsellor, Jacob Lampadius, stands out as the expert in all legal problems and the Lübeck envoy, David Gloxin, as the champion of the mercantile interest. Johann Rudolf Wettstein, of Basel, was accredited only as a representative of the Swiss Protestant cantons but acted on behalf of the whole Swiss Confederation. Lastly, the two unofficial mediators must be mentioned—the papal nuncio Fabio Chigi (later Pope Alexander VII) and, especially, the Venetian diplomat Alvise Contarini: they employed their good offices for smoothing out factional divisions and keeping alive the mutual interests of the European comity of nations.

The territorial clauses of the peace treaty all favoured France, Sweden, and their allies.

Sweden obtained the largest share: Hither Pomerania (Vorpommern), with Stettin and sole control over the Oder estuary; the Mecklenburg port of Wismar; and the archbishopric (but not the city) of Bremen and the bishopric of Verden, with control over the Elbe and Weser estuaries. In addition, there was the "satisfaction" of the Swedish Army, amounting to 5,000,000 talers guaranteed collectively by the members of the empire.

French gains

France incorporated the cities and bishoprics of Metz, Toul, and Verdun in Lorraine (French protectorates since 1552) and added to them the suzerainty over the secular vassals of the three bishops. France also obtained the ill-defined feudal rights exercised by the emperor over various towns, villages, and districts in Alsace (excluding Strassburg but including the bridgehead of Breisach), and the permanent right to garrison Philippsburg, on the right bank of the Rhine south of Speyer. These gains appeared modest. The possession of Breisach and Philippsburg, however, laid all southern Germany open to French arms; and the deliberate vagueness of the clauses relating to the cessions in Lorraine and Alsace later provided the legal or casuistical pretexts for the wars of aggression waged by Louis XIV.

The only major setback that France suffered was the conclusion of the separate peace between Spain and the United Provinces. This deprived France of an ally in the rear of the Spanish Netherlands and prolonged the Franco-Spanish War by a decade.

The gains and losses of the German princes were determined by the convenience of the principal powers: France, Sweden, and Austria. These used the claims of their lesser partners as pawns mainly for adjusting differences among themselves at somebody else's expense.

Brandenburg obtained Farther Pomerania (Hinterpommern); the bishoprics of Kammin, Halberstadt, and Minden; the county of Hohnstein; and the reversion of the archbishop of Magdeburg on the death of the existing administrator, Augustus of Saxony (which occurred in 1680). Bavaria was to keep the Upper Palatinate and the electoral dignity, but the Rhenish Palatinate was restored to Frederick V's heir, Charles Louis, for whom a new electorate was created. Saxony retained what had been secured at the Peace of Prague, namely, Lusatia and Magdeburg, but the latter only for the lifetime of the present administrator. Hesse-Kassel ousted Hesse-Darmstadt from the district of Marburg and incorporated the abbey of Hersfeld and the county of Schaumburg. The Welfs of Brunswick obtained the right to have one of their princes elected as Protestant administrator of Osnabrück in alternation with a Catholic bishop. Mecklenburg was compensated for the loss of Wismar by the bishoprics of Schwerin and Ratzeburg.

Finally, the United Provinces of the Netherlands and the Swiss Confederation were released from their legal obligations toward the empire and so recognized as independent republics.

For Germany, the Peace of Westphalia brought to a conclusion the century-old struggle between the monarchical tendencies of the emperor and the federalistic aspirations of the princes. In the course of this contest Ferdinand II had made considerable progress in revitalizing the imperial power: without consulting the electors or princes, he had arrogated to himself the outlawry of Frederick V of the Palatinate and the transfer of the latter's electoral dignity to Maximilian of Bavaria (1623) and the deposition of the dukes of Mecklenburg (1628); on his own authority, he had interpreted, that is, in fact, rescinded, the Peace of Augsburg by the Edict of Restitution (1629), thereby superseding the legislative authority of the Imperial Diet; and by the Peace of Prague (1635), he had transformed all princely troops into contingents of an army under his supreme command and abolished the princes' traditional right to conclude alliances among themselves or with foreign powers.

The Peace of Westphalia completely reversed this trend. It confirmed the full sovereignty of the members of the empire, including their right to form alliances, restricted only by the meaningless proviso "except against the emperor and *Reich*." It bound the emperor to the decisions of the Imperial Diet in all matters concerning war and peace. The Diet, which during the past 40 years had met only three times (1608, 1613, 1640–41), increased its sphere of competence at the expense of the emperor as well as the electors; the Protestant administrators of the secularized bishoprics were admitted with full voting rights. From 1663 the Diet remained in permanent session at Regensburg.

The peace established equality of rights between Catholics, Lutherans, and Calvinists: the supreme court of the empire was to be staffed by 26 Catholics and 24 Protestants, and Protestants were admitted to the Aulic Council in Vienna. The Edict of Restitution was repealed, and 1624 was declared the "standard year" according to which territories should be deemed to be in Catholic or Protestant possession. Except in the Habsburg dominions, where toleration was not granted to non-Catholics, dissidents were allowed private worship, liberty of conscience, and the right of emigration. Religious disputes were not to be decided by majority vote of the Imperial Diet but to be solved amicably between the Corpus Evangelicorum (the Protestant states collectively) and the Corpus Catholicorum, orga-

nized under the directorate of Saxony and of Mainz respectively.

Aftermath. The conclusion of a separate peace with the United Provinces helped Spain to continue the struggle against France; and Condé's victory over the Spaniards at Lens (August 20, 1648) was offset by the outbreak of the Fronde. During these civil wars in France (1649–53), the rebel leaders, including Condé, even made treaties of their own with Spain. Success came to the French from 1655, especially after the outbreak of hostilities between Spain and England and the conclusion of Anglo-French treaties of friendship (September 5, 1656) and of alliance (March 23 and May 9, 1657). Robert Blake's exploits in the Mediterranean and in the Caribbean (1655–57), followed on land with the Battle of the Dunes (June 14, 1658) and the capture of Dunkerque (June 25) and Gravelines (August 24, 1658), broke Spain's resistance. In the Franco-Spanish Treaty of the Pyrenees (November 7, 1659), France obtained Roussillon and northern Cerdagne (thus establishing the frontier along the line of the Pyrenees), Artois, and a number of frontier fortresses in the Netherlands and Luxembourg. Spain had ceded to France the first place among the great powers of Europe.

Treaty of the Pyrenees

The last decade of the 50-year period here surveyed witnessed another conflict in northern Europe. It arose, however, from the start of the Thirteen Years' War between Russia and Poland (1654–67), which is best considered as belonging to a later historical period, characterized by the decline of Poland and by the rise of Russia and of Brandenburg-Prussia. Insofar as it reaffirmed certain tendencies of the earlier period, it may be briefly summarized here.

The question of the Ukraine led to war between Russia and Poland in 1654. Charles X of Sweden took advantage of this war to attack Poland in July 1655. The elector Frederick William of Brandenburg took the Swedish side from 1656 to 1657 but changed to the Polish in 1658. Russia began war against the Swedes in June 1656 but concluded the Truce of Valiesari in December 1658. In 1657 the emperor Ferdinand III's heir, Leopold I, allied Austria with Poland. Frederick III of Denmark attacked Sweden in June 1657.

Baltic settlement

By the Peace of Roskilde (February 26, 1658), Denmark was forced to cede to Sweden not only Trondheim on the North Sea coast of Norway but also Bohuslän on the Swedish coast at the end of the Skagerrak, Skåne on the eastern side of the Sound, with adjacent Blekinge, and the island of Bornholm, thus forfeiting all hope of the "lordship of the Baltic." Denmark, however, subsequently refused to close the Baltic to Western shipping and was attacked by Sweden again in summer 1658. The Dutch, in their own commercial interests, came to Denmark's support, and the Treaty of Copenhagen (June 6, 1660) restored Bornholm and also Trondheim to Denmark, but otherwise confirmed the settlement of Roskilde.

The Peace of Oliva (May 3, 1660), between Sweden on the one hand and Poland, Austria, and Brandenburg on the other, assigned Livonia to Sweden and recognized Brandenburg's full sovereignty over ducal Prussia, both these stipulations being made at Poland's expense. The Russo-Swedish Peace of Kardis (July 1, 1661) reestablished the terms of the Peace of Stolbova (1617).

France had played an active role in mediating the Baltic settlement. Habsburg Austria had gained nothing from the conflict and was meanwhile confronted by the League of the Rhine (August 14, 1658), organized by Mazarin, whereby France, Sweden, the electoral archbishoprics (Mainz, Cologne, and Trier), Münster, Brunswick-Lüneburg, Palatinate-Neuburg, and Hesse-Kassel guaranteed the Peace of Westphalia against Habsburg revisionism.

(S.H.St./Ed.)

ECONOMIC DEVELOPMENT OF THE EARLY MODERN WORLD

The decline of the feudal system and the growth of commerce. No sharp divide separates the economic life of the Middle Ages from that of the early modern world. But, in retrospect, much of the 15th century in Europe displayed the symptoms of the decay of an old order. Population in many places was smaller than a century and a half earlier, and prices had generally declined. These changes were reflected in both the rural and the urban economies.

In those areas of Europe where a manorial type of agrarian economy existed, the links between lord and peasant were becoming either looser or tighter. The customary tenant, who held his land in return for services performed on the lord's demesne (estate), was in some places becoming a rent-paying tenant, even (in England, for example) turning into a tenant-farmer exploiting the disintegrating demesne land. In eastern Europe, beyond the Elbe River, however, the reverse was happening. There the great lords were tightening their hold on the peasants' lands. Peasant was becoming serf and serf becoming slave. Thus, the old balance within feudalism was being tipped one way or the other.

Trade and industry were similarly changing in character. The traditional organization of the city crafts into guilds was breaking up, and companies of merchants, on the one hand, and of journeymen, on the other, began to assume responsibility for the specialized functions of their members. Germany in the 15th century suffered continual urban conflicts as the guilds struggled among themselves for power and privilege, and everywhere in Europe guild control over industry receded in the face of an almost universal tendency for individual urban capitalists, or groups of capitalists, to assume control, through mercantile capital, over the entire process of manufacture in textiles, leatherworking, mining, and the metallurgical industry.

The structure and direction of international trade was also changing. The cities of the Hanseatic League (a league of towns in northern Germany that banded together in the latter part of the Middle Ages to protect and develop their foreign trade), which had long been dominant economically and politically in northeast Europe, had fallen into fratricidal war. Their conflict with the Scandinavian powers in the 16th century virtually brought about their final collapse, and this, in turn, had a depressing effect on those important sectors of economic life in such cities as Bruges, in Flanders, and London, which had been closely dependent upon the League. North Italy, which had long been the most advanced economic area, was also under severe pressure. The four great cities of Venice, Milan, Genoa, and Florence, which had dominated the life of the Po Valley, the Tuscan Plain, and large areas of the Mediterranean, were now faced with new competition and problems. The great voyages of discovery had opened up by 1500 an alternative route to the Orient around the southern tip of Africa, thus threatening the overland spice trade of Venice, although the effects of this were slow to be felt and their incidence was only intermittent until well into the 17th century. More serious was a slow but steady shift of the centre of European economic activity away from the Mediterranean toward the Low Countries. North Italy still remained an advanced area of wool and silk production and glass and porcelain manufacture and a centre for the great merchants in spices, grain, alum, wine, sugar, and other exotic products. Venice and Genoa also showed a sinuous flexibility in meeting the competition of the Portuguese spice traders. But the 16th century was to be the age of Antwerp, in the Netherlands, rather than of Venice, and its successor in the 17th century was a north Netherlands economy dominating the world from the great entrepôt of Amsterdam.

The Hanseatic League

The influence of population on economic growth. Modern research has established almost beyond doubt that behind the new impetus stimulating the developments of the 16th century lies the reversal of the population decline of the 15th century. In its last decades there began the steady growth that may be detected everywhere in Europe throughout the 16th century and seems to level off only by the middle of the 17th century. By 1600 Europe had a population approaching about 100,000,000. Of these, perhaps one-half lived in countries bordering the Mediterranean. This probably represented between 20 and 30 percent more than in 1450 and was to increase by another 30 or 40 percent by the middle of the 18th century. Most countries shared in the 16th-century increase, especially

Italy, England, the Low Countries, and France. By 1700 it was apparent that the increase was slowing down and even declining in some countries. The population of Germany was severely cut by the disease and poverty that followed the Thirty Years' War, and the population of Spain may have dropped from 8,000,000 to 6,000,000 in the 17th century. But, in general, the European economy reflected the buoyant increases in areas such as England and the Low Countries, especially in the area that today constitutes The Netherlands, where the population virtually doubled between 1500 and 1650.

The reasons for this increase have been much debated. Population movement must depend upon the balance between the birth rate and the death rate. Mortality, especially through the bubonic plague, which attacked everywhere sporadically, was extraordinarily high, but the incidence of plague seems to have slackened in the 16th century, and it finally disappeared from western Europe by the beginning of the 18th century. The declining virulence of the plague, coupled with the removal or relaxation of restraints on early marriage, thereby facilitated a rise in the birth rate. This, in turn, was followed by urban growth, the decline of guild and apprenticeship laws, and the increase of individual farm holdings.

Bubonic plague

If the origins of this demographic revolution are obscure, its consequences, on the contrary, are quite clear. Everywhere it generated an increased demand for food and the necessities of life. This increase, in turn, not only became a major factor in the great inflation of the 16th century but also brought into being a new class of entrepreneurs, or middlemen, who seized the opportunities for trade and profit that it offered.

The rise of Antwerp and Amsterdam. The rise of Antwerp in the first half of the 16th century to the position of the leading entrepôt in the world vividly reflected this new movement. From about 15,000 in the mid-15th century, its population had passed the 100,000 mark by the middle of the 16th century. Antwerp was supreme during the half century from 1520 to 1576 as the warehouse and market of world trade. It not only absorbed the industrial production of its own textile centres, such as Ghent, Lille, and Ypres, but it also formed a sophisticated centre for merchant adventurers, bankers, and speculators from all over Europe. The merchant adventurers from England concentrated the largest part of the export trade in English cloth in Antwerp. The south German bankers (Fuggers, Welsers, Paumgartners) traded in cloth, spices, and metals with Germany and Italy. Italian merchants sold the produce of north Italy into northern Europe through the Antwerp market, while agents for the kings of Portugal sold great cargoes of Indian pepper, nutmeg, cinnamon, and cloves. Ships from Cádiz brought in cargoes of Spanish wool and wine and, later, of silver, to pay for purchases of cloth, iron, coal, and glass or to repay the financiers who fed, armed, and clothed the Spanish and Imperial armies.

Antwerp was not only a centre of trade but also of finance, where the European governments borrowed on a vast scale, especially for war purposes. In England the Tudor financial agent Sir Thomas Gresham acted as royal factor from 1551 to 1574, with the duties of raising long- and short-term loans for the crown on the Antwerp money market. Most other kings and princes did the same. Among the great banking houses were those originating from south Germany, of whom the Fuggers were the greatest. Starting as weavers and manufacturers of cheap textiles in Augsburg in the late 14th century, the Fuggers had risen to be a world trading and banking house. They were now among the leading exponents of the arts of public finance, farming the revenues of the emperor and advancing capital against concessions of silver, copper, and iron mines in the Habsburg territories. They were thus the greatest subsidizers of European war and dynasticism, lending 500,000 florins to secure the election of Charles V as Holy Roman Emperor in 1519.

The Fugger banking houses

The sack of Antwerp in 1576 by Spanish troops added severely to the economic problems that beset Antwerp increasingly after the 1550s. Its capture by Alessandro Farnese, duke of Parma, in 1585 sealed its fate, for the Dutch blocked the approaches of the Scheldt River and virtually ended Antwerp's career as a world market for two and a half centuries.

The stream of refugees from the southern Netherlands began in the 1550s and continued for half a century but was at its peak in the 1570s and 1580s. Up to this time the northern Netherlands were much poorer than their southern neighbour, but, with the aid of the migrants, much wealth, enterprise, and skill was transferred across the great rivers from the vulnerable south to the more defensible north. Amsterdam, seceding from Spanish loyalty to the rebels in 1578, swiftly succeeded to the position lost by Antwerp. At Leiden immigrant merchants and textile workers from the area around Ypres launched a variety of woollen industries making the "new draperies." At Haarlem others set up bleaching centres that attracted great quantities of coarse linens from Germany. Other southern industries, such as silk weaving, glass blowing, paper making, sugar boiling, distilling, and printing, also moved to Amsterdam.

The growth of international trade. The great migration out of the southern Netherlands exported much talent and wealth also to London, Norwich, Colchester, and other English towns and cities, bringing the new textile skills and much mercantile and financial technique to England. But at this stage the Dutch republic was the greatest gainer, and between 1585 and 1620 Amsterdam became the centre of a world network of trade. This network was extended until it reached from the Baltic to the Mediterranean, from the British Isles deep into Germany, and throughout the 17th century it was increasingly linked to India and the East by the operations of the new Dutch East India Company (1602) and to America, the West Indies, and Brazil through the Dutch West India Company (1621). As a victualling station, the southern tip of Africa was settled with Dutch farmers, while Dutch slave traders built castles on the coast of West Africa. As merchants, financiers, shippers, and drainage engineers, the Dutch were as ubiquitous in the 17th century as were the Scots in the 19th.

Dutch East India Company

The Bank of Amsterdam (1609), the Loan Bank (1614), the Bourse (1609), the Marine Assurance Chamber (1598), and the Grain Bourse (1616) collectively formed the central machinery of an economy through which a large proportion of world trade passed. Except for the great herring fishery and the cloth industry of Leiden (the largest in the world), the Dutch were middlemen, carriers, and brokers rather than industrial producers, but their commercial skill, advanced financial techniques (including a fully developed stock exchange and company-share system), together with a low rate of interest, abundant capital, and a willingness to put it at risk overseas as well as at home, gave them a commanding position in the 17th-century world economy. The silver that represented the profit won in their trade with Spain was not hoarded but used to purchase goods from areas such as the East Indies and the Baltic, and these, in turn, were resold to the European nations, including Spain itself. Above all, the Dutch were the first to develop a purpose-built, oceangoing boat (the "flyboat"), which was unarmed and cheap to build and operate. With the cheap freight rates provided by the flyboat, the Dutch captured such great bulk trades as corn, timber, salt, and sugar.

Repression in the agricultural sector. The demand for Baltic grain, arising from the general increase in European population, which had everywhere turned food surpluses into deficits, was the greatest commercial opportunity seized by the Dutch. Throughout the 17th century a procession of Dutch flyboats passed through the Danish Sound carrying cargoes of grain from the southern shores of the Baltic Sea. To supply this demand, the large-scale farming system, based everywhere on the process that came to be known as *Bauernlegen* ("repressing the peasants"), spread throughout central and eastern Europe. Throughout eastern Germany, Austria, Hungary, Bohemia, Poland, and Russia, a process of increasing serfdom set in, culminating by the mid-18th century in Russia, where peasants were bought and sold like cattle. The rural population repressed thus had actually enjoyed more freedom than their counterparts in medieval western Europe, but now

Increase in serfdom

the great demesnes grew as prices rose and corn exports increased. Germans and Slavs were lumped together in virtual slavery in Mecklenburg and Pomerania. In Bohemia the German Catholic conquerors imposed brutal terms on the Czech heretic population. Conditions did not improve until the early 18th century in central Europe, and in Russia serfdom survived until the 19th century. Everywhere, the system produced economic stagnation, depriving trade and industry of capital and labour.

The agrarian system around the Mediterranean was less harsh, but, even there, the scarcity of fertile land, such as the Po Valley, the Roman Campagna, and the Tuscan Plain, gave rise to severe social problems. In Italy the limited areas of cultivable land demanded heavy capital investment, either for irrigation or drainage. The nobility or the municipality that provided the capital and carried out such works naturally demanded the lion's share of the profit. Hence the bitter comment of a 16th-century writer: "The plains belong to the lord, the mountains belong to the peasant."

Yet, sun, bread, and wine were not lacking. Rice and corn were among the crops that the enterprise of Venetian and Milanese landowners added to the Mediterranean scene. Italy therefore stood halfway between the stagnation of the East and the modest progress of the West, where some increase in productivity was discernible. England, northern France, and western Germany formed an area quite different from that in the East. In the West lords tended to become *rentiers*. Peasants redeemed their quitrents; in France many gained the status of something like owner-occupiers, and in England many became tenant farmers. Others declined into landless labourers. Providence offered a variety of fates; but all were free men, and the society of which they were a part could move on to a commercial or industrial future, exchanging feudalism for a reformed monarchy or constitutionalism.

The consequences of the different agrarian economic systems and social structures adopted east and west of the Elbe were very marked. The social organization and capital investment in Western countries, together with modest improvements in plows, implements, drainage, manure, and enclosure, resulted in higher yields and a perceptible increase in productivity, especially by the second half of the 18th century. In east Europe, serfdom and stagnation produced an actual decline in productivity from 1650 onward.

The rise of the entrepreneur and the labour market. The 16th century was an age of industrial expansion that in some places, such as Britain, continued into the second half of the 17th century. Elsewhere, it levelled off by the middle of the 17th century, as, for example, in Holland. In Germany, progress was halted by the Thirty Years' War. This age of expansion resulted from many related causes: price inflation, population growth, the demands of governments in the new centralized states for war and court luxury (France, Spain, Sweden, and Britain).

Increased demand led to greater production. Since industry in this period most often meant production of consumer necessities, it relied for its raw materials everywhere on the countryside. Agriculture produced the wool, flax, and leather that formed the basis of the clothing industries supplying rich and poor alike. It also provided the corn for the baker and barley for the brewer. Brewing was one of the few great capitalist industries employing intensive capital equipment in the centuries before the Industrial Revolution, and the great brewers of London and the large cities of Holland and Germany were among the largest contemporary capitalists. Vineyards supplied the wine industries of France, Italy, Spain, and Germany. In addition, princes and noble landowners owned the rights over minerals that lay under the soil of their estates. They, therefore, exploited their mines and the deposits themselves or (more commonly) sold concessions to urban capitalists to do it for them.

Brewing

Role of the entrepreneurs. Increasing demand raised the incomes not only of urban entrepreneurs, who seized the opportunities offered by new demands, but also of those noblemen and gentry who, by luck or shrewdness, extracted opportunity from their estates. Throughout the enormously varied mining and metallurgical enterprises of western and eastern Europe (including coal, iron, tin, silver, copper, lead, and alum) and the glass and ceramics, chemical, textile, paper-making, and many other industries, there is clear evidence of the increasing role played by the capitalist organizers. These consisted mainly of city merchants whose capital was increasingly required to finance growing output and new technology. Such changes are most clearly reflected in the growth of these merchants in cities like London; Antwerp; Amsterdam; Cologne, Augsburg, and Nürnberg (Germany); Lyons (France); and, to a lesser extent, in smaller cities such as Norwich (England), Leiden, Haarlem, Hamburg, Barcelona, and Milan. In the Dutch cities they formed after a generation or two a merchant elite influencing the policies of the state at home and abroad. The philosophy of the Dutch republic is most vividly and accurately represented in *Het Interest van Holland* (1662; *Political Maxims of the State of Holland*), allegedly by Johan de Witt. In reality, this was the work of a great Leiden textile entrepreneur, Pieter de la Court. Its theme was the need for a context of peace, neutrality, tolerance, and economic freedom in which trade and industry could flourish. In London, a representative figure earlier in the century was Alderman Sir William Cokayne, confidant of the King and the contriver of 12 great commercial and industrial schemes, including a project for modernizing and expanding English textile manufacture. This came hopelessly to grief, however, and precipitated one of the greatest crises of the age by 1618. Another entrepreneur was Sir Hugh Myddleton, brother of a lord mayor of London, a mining contractor of Welsh origin with great enterprises in Wales and founder of the New River Company, which between 1609 and 1620 dug a freshwater channel 40 miles (60 kilometres) long to supply London with drinking water. In Europe the south German bankers performed similar functions as financiers and organizers for German industry. Dutch capitalists originating from Liège similarly exploited the iron and copper resources of Sweden. The Trips and de Geers were the most prominent of a large group of Netherlanders who settled in Scandinavia as agents of the monarchy to develop mining and metallurgy.

Early industrial schemes in London

Use of capital. The accumulated wealth of these new merchant princes was either saved and loaned as capital to borrowers, including the state, traders, and industrialists, or spent on personal consumption. In England a good deal of their wealth was spent on social advancement. Cokayne, for example, succeeded in getting all his daughters married to peers. Great merchants commonly moved out of the city and bought country estates for themselves in the counties surrounding London. Dutch merchants, though tending to remain in trade, also built themselves country palaces on the Vecht River near Amsterdam, just as Venetian merchants built great mansions along the Brenta River. Both invested capital in the countryside, purchasing farms and undertaking drainage and general improvement schemes. In the middle of the 17th century, Amsterdam capitalists financed and implemented the drainage of great areas of the English Fenland under the eye of the great Brabant engineer Cornelius Vermuyden.

Rise of the landowner-entrepreneur. Such enterprises were not entirely limited to the urban merchant. Everywhere in Europe, landowners joined in exploiting their own mineral deposits, such as the Lowther family in northwest England and the Dudleys in the English Midlands. At Stolberg in Germany, the counts encouraged mining and iron working on their estates, while Duke Julius did much to organize and modernize mining in the Ober Harz. One of the most famous woollen mills in central Europe was that of Count Waldstein at Oberleutensdorf. Many other nobles used the forced labour of peasants to develop textile production in Bohemia and Moravia. Others, in Silesia and Bohemia, exploited clay deposits on their estates to develop famous porcelain manufactures. Many a princely palace had a *Porzellan-zimmer* ("porcelain room") exhibiting products from the estate. Especially in England, but in many other countries also, there was intermarriage between members of these commercially oriented noble families and their aspiring bourgeois partners.

Technological innovations and improvements. The early modern period was not an age of revolutionary invention, but industrial expansion in some cases led to a number of improvements, mostly of an empirical kind. In mining, galleries and shafts were improved. Hoists brought up the ore and water. As early as 1550, in Bohemia, pumps worked by water power made possible shafts sunk to a depth of 1,300 feet (400 metres). Waterwheels were used to drive hammer forges and rolling mills. New methods of making brass and copper and lead alloy were discovered. Blast furnaces replaced open hearths. Coke was substituted for charcoal by 1700 to smelt lead, tin, and copper ores. Later, Shropshire ironmasters smelted iron similarly. In the Netherlands windmills were improved and used for sawing timber. Everywhere textile industries developed improved equipment: fulling mills were driven by water; the Dutch at Leiden developed a multiple ribbon loom that was borrowed for English use; William Lee, an Englishman, had invented his frame knitting machine in 1589. In spite of opposition, mechanization slowly established itself in the succeeding centuries.

Use of windmills

Need for commercial capital. These technological improvements that accompanied and made possible economic expansion had one thing in common: they cost capital; and these higher costs, mainly for working capital but also for fixed plants, equipment, tools, and buildings, explain the need for a new organization in industry and for a new type of entrepreneur to carry it on. Probably the most highly capitalized industry of all was brewing. By the end of the 16th century, the great London brewers, many of them of Dutch or German extraction, were producing beer in breweries that already bore a resemblance to a modern plant. To a lesser extent, this was also true of the great glassworks of London and the English Midlands. A late 16th-century Leipzig merchant of great wealth, Henry Cramer, had a textile factory, representing a large capital outlay, that included fulling mills, weaving sheds, a dye plant, and a complete residential area for the workers. In the Swedish iron and copper industries, capital was provided by the so-called *Brukspatroner* (proprietor), who provided the necessary credit. In England, the Netherlands, and in a rather less sophisticated manner in many centres of France, Germany, Switzerland, and Bohemia, the complicated processes needed to produce a piece of cloth—wool sorting, carding or combing, spinning, weaving, fulling, dyeing, and finishing—were organized under the control of an urban capitalist. His function was to provide credit, pay wages, set the standards and designs to be followed, and, finally, to seek markets for the product.

The spread of industrial skills. Much transfer of skill was involved in the rise of the pre-Industrial Revolution's manufacturing system. The nurseries of technology can be identified as north Italy, the Low Countries, and south Germany. The 16th century had no blueprints, so both capital and technology were transmitted across national boundaries in the most highly personalized way. Colonies of Italian merchant bankers showed Antwerp and Bruges the secrets of double-entry bookkeeping, and Italians dominated the London money market until well into the 17th century. One-third of the population of Norwich in 1600 was of Dutch origin. There, as in London, Colchester, and many manufacturing centres in the east of England and the Midlands, Dutch, Flemish, and Walloon experts were to be found in textiles, brewing, glassblowing, printing and bookbinding, horticulture, and many other occupations. The Höchstetters and other south Germans bought concessions to develop the Cumberland mines of England, as they did in Spain and throughout the territories of the Holy Roman emperor. The process continued after the expulsion of the Huguenots from France in 1685, when another wave of highly skilled merchants, artisans, and professional men flooded into the Dutch republic, England, and Prussia, where they reinforced strongly (and in some cases originated) the processes of invention and development in such industries as linen and silk manufacture, clock making, paper making, and hat making.

Italian merchant bankers

The guild system and the labour market. Thus, nobles, merchants, and, in some cases, middlement rising from the ranks of industry itself, inaugurated major changes in the early modern economy throughout Europe. Rising demands caused industry to outrun the supply of skilled labour (journeymen) controlled by the urban guilds. Therefore, the unskilled and unorganized rural workers were recruited into a new industrial system known as the "putting-out" system, or domestic system. Under this system, the capital, material, and, sometimes, tools were supplied by a merchant capitalist, who coordinated the different processes necessary to manufacture and disposed of the final product on the market. The processes (*e.g.*, combing, spinning, weaving, etc.) were carried on by a widely dispersed system of cottage labour.

This did not mean the guilds immediately disappeared. In some places, for example London, they changed their nature as the great merchant companies gained control of the craft guilds through their control of capital and power over the market. In other English towns, after a long series of amalgamations, the guilds weakened and finally disappeared. There were few left by the end of the 17th century, and industry increasingly was left to its own devices. The Industrial Revolution of the 18th century was able to spread all the more rapidly because only the vestigial remains of the old guild system were left.

In France the economic reorganizations of Jean-Baptiste Colbert revived the guild structure largely to create revenue for the crown, but even in France guild power was weakened by the rise of the putting-out system in rural areas and the grant of licenses and exemptions by the crown to individual entrepreneurs in return for a fee. In Spain guilds actually increased—a characteristic symptom of Spanish backwardness. Likewise, in Germany and central Europe, the roots of the guild system lay deep and withered away only slowly as the enlightened despots of the 18th century introduced liberty of craftsmanship. In Italy the situation varied. In Rome guilds grew in number, whereas in Tuscany local rulers subordinated guilds to themselves. In Venice their power seems to have constituted a major obstacle to technological progress, as in textiles, for example. In Poland, as in Germany and Spain, the guilds remained strong. In Russia, oddly enough, there was far less regimentation until Peter I the Great reinforced the guild system as part of his westernization program.

Survival of the guilds

Broadly, guild power was undermined most easily in the great entrepôt ports of northwest Europe, where capital and enterprise were relatively plentiful and where sea and water transport enabled economic change to come about most easily. The putting-out system, which had existed here and there in the medieval economy, now spread to many branches of production—such as mining, metallurgy, textiles, printing, and paper making. These were all industries demanding capital for equipment, raw materials, the payment of a large labour force, and a knowledge of marketing. Under the new system, guild and free labour might be combined. Spinning and weaving could be done in the countryside; combing and dyeing, in the towns. Thus, hierarchies of workmen grew up appropriate to each individual industry. In the metal industry of Solingen (Germany) the swordsmith, the temperer, and the sword furbisher all took part in the manufacture of a sword, but only the last (the *Reider*) was allowed to market the product. The same tendency for the man who was nearest to the market to assume control of the industry can be seen throughout the Birmingham and Sheffield industries making small ironware.

Emergence of large-scale enterprises. Large-scale production was not typical of the 16th century. Here and there an entrepreneur might concentrate production under one roof, as the famous Tudor manufacturer Jack of Newbury (John Winchcombe) is said to have done, employing some 500 or 600 adults and another 400 or 500 children. Less spectacular instances were to be found in late medieval Florence and Flanders. In the late 17th century such examples multiplied not only in western Europe but also in Bohemia and Russia. Although such examples are exceptions to the general rule of a small unit of production, they show how industry was everywhere trying to burst out of its medieval straightjacket and assume the forms of organization most appropriate to the relatively economic and efficient use of capital, labour, and technology. Trans-

Factors influencing industrial locations

port problems tended to restrict industrial location to the places where raw materials, labour, fuel, and power were available, but, where goods could be carried by sea or canal or where workers could be attracted by better pay or conditions of life, industries grew up despite such locational restrictions. The desire of governments for greater income or more power likewise reinforced the tendency to set up industries in places not naturally suitable for them.

Unemployment and social problems. This new economy, larger but more flexible than its predecessors, provided greater opportunities for employment but had greater social consequences. The estimates of Gregory King, the late 17th-century English social statistician, suggest that at least one-third of the English population was unemployed or underemployed. This phenomenon was remarked on increasingly by many observers in England, Spain, Germany, and Italy in the 17th century. It accounted for the attempt to gather the poor together in workhouses that were supposed to combine the functions of poor relief and larger national production. Much criticized and only fitfully effective, it served as a model for the more efficient deployment of resources by private enterprise in factories that was to characterize the Industrial Revolution itself.

Capitalism and the Protestant Ethic. Historians and sociologists who have looked for the deeper motivations of 16th- and 17th-century capitalism have sometimes suggested that they may be found in Protestantism, especially Puritanism. In 1904 the German sociologist Max Weber put forward the proposition that the linked phenomena of capitalism, laissez-faire, and the justification of business as a "calling" all derived their warrant from these sources. In 1922, Richard Henry Tawney, an English economic historian, tried to substantiate this thesis. Calvinism, in particular, was designed for a rising bourgeois class that was "conscious of the contrast between its own standards and those of a laxer world, proud of its vocation as the standard-bearer of the economic virtues" (*Religion and the Rise of Capitalism* [1926]). In spite of counter-arguments, such theories have had wide influence. Yet, they now need to be re-examined in the light of the knowledge that has emerged since they were first put forward. This includes the realization that the growth of capitalism in the early modern world received many stimuli unknown to Weber and Tawney, including the growth of population. Some of the changes that were supposed to be specifically linked with north-European Protestantism also involved no real innovations, being rather a shift to northern Europe of economic practices well established in north Italy, the cradle of the modern European economic system.

The Weberian thesis

Calvinism. Further examination of the basic theology of Martin Luther and John Calvin suggests a medieval rather than a modern economic outlook, containing little to appeal to a rising bourgeoisie. Even Calvin gave his views on problems of economic morality with caution and reluctance. They occupy only a few pages in his 20 volumes of printed works. What Calvin did was to remove the question of usury from the complex of theological pedantry and place it on a basis of simple morality. Reasonable interest on a loan was no more unjust than a reasonable rent for land, but extortionate interest squeezed from the poor was "evil and foul." His disciples took a similar line. (The doyen of Dutch Calvinist theologians, Gisbertus Voetius, excommunicated a God-fearing woman in his congregation because her husband was a pawnbroker.) Generally, Protestant countries adopted attitudes to these problems very similar to the Catholic tradition, fixing maximum rates of interest, controlling prices, and setting up public pawnshops.

The Puritan societies of such areas as the Netherlands, East Anglia in England, New England in North America, and Huguenot parts of France likewise contradict the assumption that Calvinism was particularly associated with merchant capitalism. The Protestants who fled from the southern Netherlands before Spanish persecution into Holland and eastern England seemed to have been mainly skilled artisans. The capitalists among them, however, constituted only a small minority of the whole. Calvinist influence is clearly discernible among the personnel and policies of the Dutch West India Company, but, in general, dogmatic Calvinism was not characteristic of the rich merchant elite of the Netherlands. They were, for the most part, liberal and anticlerical in outlook, no more inclined to trust dynasticism or theocracy in Protestant form than in Catholic form. Indeed, in the 17th century, Dutch politics turned largely on the conflict between these "libertine" merchant princes of the Dutch republic and the party known as Orangists, led by the Calvinist ministers but composed largely of the humbler social orders in support of the House of Orange. The Orangists wanted to continue the religious crusade against Spain, which the great merchants realized would be disastrous to their business.

Protestant refugees from the Netherlands

The settlement in New England in North America was theologically Puritan and socially medieval. It was not until the middle of the 17th century that newer immigrants from the City of London began to challenge the older ideals of social stability, order, and discipline inherited from the Pilgrim Fathers. Like their London or Amsterdam contemporaries, these new capitalist elements wanted mobility, consumption, even luxury, all thoroughly deplored by men of the old dispensation, such as Gov. John Winthrop.

Similar conflicts may be found among the Huguenots of the areas around Bordeaux and Marseilles in France. Probably the majority of Huguenot support came, here as elsewhere, from the lower orders together with the lesser gentry, both hard hit by rising prices. Rich merchants may certainly be found, but they were probably no more important as an element than professional lawyers or doctors.

Origins of the Weber thesis. Such observations throw considerable doubt on the Weber thesis, which may have received some of its plausibility from later evidence of the predominance of Dissenters in 18th- and 19th-century British economic life. But much of this was Quaker or Methodist and based on a theology totally different from that of Calvin. It may well have been the result of the exclusion of Dissenters from many spheres of public life after 1660. In general, it looks as if the majority of the adherents to radical forms of Protestantism came from the ranks of displaced persons from the lower orders of society. The new economic developments tended to take place beyond the control of the old medieval towns and guilds. In what is now Belgium and, later, in Holland, there was a floating population from the ports and the countryside, much at the mercy of the economic cycles of alternating prosperity and depression. Religious persecution was another force turning thousands of artisans adrift in search of new places of settlement. The enforced mobility and restlessness of the times seemed to have given a special attraction to class, status, or wealth. Radical religions that emphasized the importance of character and conduct and that even preached socially egalitarian doctrines seemed to give purpose even to humble occupations. The doctrines of election and predestination had a special appeal to those who could believe that economic success based on thrift and sobriety was clear evidence of God's favour. It seems probable that in the casual, easy-going society of the early modern world, men thus motivated toward steady and methodical work would inevitably tend to rise in the social scale. On this interpretation, while the initial attraction of the new religions to capitalists may have been exaggerated, the practice of their beliefs and social habits may have contributed to the rapid growth of a class of capitalists later on. At the upper end of the social scale also, it has been argued that the appeal of Calvinism was not so much to the prosperous gentry as to those who were in difficulties through price inflation.

The Dissenters

The voyages of discovery and the price revolution. *The development of chartered companies.* The late 16th century saw the nations of western Europe—Spain, Portugal, England, and the emergent Dutch republic—locked in the first scramble for trade and empire. Some of the contestants were individual merchants who attempted to conduct their business in the new areas in the traditional way, either as individual adventurers or at most loosely associated with other merchants in a fellowship or society. This was the way the merchant adventurers had conducted a large part of England's export trade in cloth to northwest Europe. It was an association of individuals in which every member

traded alone with his own stock and employees, subject only to the regulations imposed by the company in order to secure benefits for all members. This was what was called the "regulated company," and it did not represent any accumulation of common capital. Such an institution had serious defects when it came to organizing and financing the much longer voyages, employing larger armed ships with larger crews and with larger cargoes at stake. Voyages to the new areas of discovery often necessitated heavy outlays on defensive forts and equipment to protect the adventurers against hostile attacks. Such ventures were more than an individual investor could finance, and they required the collection of capital from many investors and its professional management. Such methods had occasionally been employed in the Middle Ages. These joint-stock methods now became far more common. In the second half of the 16th century, London capitalists, aristocratic and mercantile, raised money jointly to finance Sir John Hawkins' slave journeys, Sir Martin Frobisher's search for the Northwest Passage, Sir Francis Drake's raids on Spanish America, Sir Walter Raleigh's colonization of North America, and the trade enterprises to Russia and the Levant. Similar methods were followed by the adventurers of France and Holland, but very few of them succeeded.

The enterprises most significant for the immediate future were those to the East Indies. In Holland 12 groups of investors in the 1590s joined to finance such voyages by a separate parcel of capital, but the profits were sadly reduced by the competitive buying and selling of the valuable spices by the rival entrepreneurs. In 1602, therefore, they were amalgamated into the United East India Company. Like most Dutch institutions, this was a federation—six "chambers" situated in six different towns or provinces, each with its own organization and capital and subject only to general control from the central body. The need for permanent fixed capital for forts, ships, and arms suggested that the capital should also be permanent. A stockholder who wanted to get his money back could only do so by selling his claim on a share of the capital to somebody else at the price it would fetch on the market at the time of selling. Thus, the East India Company shares became one of the first and largest blocks of share capital to be traded on the new Amsterdam Bourse.

The United East India Company

Anglo-Dutch rivalry. The English East India Company obtained its charter on the last day of 1600. A company of 220 adventurers was established, which sounded very like a regulated company, but they recognized that India was too far away to be exploited by individual enterprise. Therefore, there was to be "a joint and united stock," though each voyage remained a separate capital investment. It was more than 50 years before the English adopted the Dutch method of a permanent capital fund.

The Dutch were the pacemakers under the vigorous, ruthless leadership of Jan Pieterszoon Coen, governor in the East Indies until 1629. By 1680 they had a monopoly of the spice-producing regions of the Malay Peninsula and Archipelago. They held Ceylon and many points on the mainland of India, the Persian Gulf, the Red Sea, China, and Japan. They employed every device to maximize profits on the cargoes of Japanese copper, silk, and porcelain; Chinese tea, pottery, and wallpaper; Indian cottons and spices; Persian carpets; Arabian coffee; Malayan tin; but, above all, the precious spices of the Malayan Islands, which left Asia at Christmas each year.

The English company was less successful. It managed to make high profits on the early voyages and set up many trading posts in the East, but it was evicted by Japan and driven out of the Spice Islands by the Dutch. On the Indian mainland the costs of defense were high. Lacking the expertise and the resources that enabled the Dutch to organize a self-balancing trade in many Eastern areas, the English company was always under pressure to export much larger quantities of silver to pay for its Eastern purchases than the Dutch. Throughout the 17th and 18th centuries it faced constant criticism for draining England of silver. Its reply was that it resold a large proportion of the spices, textiles, tea, coffee, and porcelain that it brought back from India in other European markets, thereby winning a net surplus of treasure.

The Dutch were also active in North America, where colonists had settled in the New Netherlands, along the Hudson River, trading with Newfoundland, Virginia, and the West Indies. Their hold was not strong, and it was easily broken by the English in 1664. Nor were their enterprises in the West Indies and South America completely successful. The Dutch West India Company was formed as much to fight Spain as to develop trade. Large areas of Brazil were annexed but lost again to Portugal. The sugar market was a dangerously fluctuating one. Dutch Guiana, with its coffee, cacao, and cotton, produced better results, as did the slaving ports on the coast of West Africa, which were a main source of supply to America and the West Indies. But the West India Company was not, in the main, successful.

North American trade

Success and failure of the companies. Indeed, many of the European chartered companies suffered results varying from barely successful to disastrous. The Virginia Company was a costly failure. The 30 or more French companies set up in the 17th century, including an East India, West India, and Senegal, did little but consume French capital. The English Royal African Company had a checkered career, as did competing companies formed in Denmark and Prussia to exploit Africa and the East Indies.

Somewhat late in the field, France managed to retrieve an otherwise dismal string of failures in the Caribbean. In spite of the collapse of Colbert's West India Company, individual French traders did well in what was known as the Sugar Bowl, developing sugar plantations on Martinique, Guadeloupe, Haiti, and other smaller Caribbean islands.

By 1700 a century of company activity by the Western nations in America, India, Asia, and Africa had left indelible marks. South America and the West Indies had been designated forever as "Latin" America. With the Dutch trading stations eliminated, North America's future was poised between England and France. England was firmly entrenched on the Indian subcontinent, while to the East the Dutch were equally the masters of the Malay Peninsula and Archipelago, with regular trade connections to China and Japan. It is likely that the Dutch East India Company had the best record of profit. None of the other company enterprises had anything but a patchy record, and some had lost enormous sums of capital and seen little, if any, profit. Some individual merchants in Amsterdam, Middelburg (Holland), Rouen and Nantes (France), London, and Bristol (England) had made personal fortunes. Others had lost them.

Success of the Dutch East India Company

But, if the colonial trades were by no means the easy gold rush they have sometimes been thought, it must be remembered that the capital investment they represented was still only a small part of the total trade of the Western world. The bulk of trade and shipping was still inter-European. The spices, sugar, cotton, and other goods were commodities that leavened the lump, but they were far from being the dominant influence.

The influx of silver and price inflation. One commodity produced in the New World nevertheless had a direct bearing upon an economic movement that, in turn, had incalculable effects upon the economy and society of the 16th and 17th centuries. The relationship between the flow of Spanish-American silver and the price inflation of the 16th century is still a matter of debate, but many historians still maintain that the major cause of the price inflation lies in this addition to the monetary supply of the world; and most others would agree that it played at least an important role. The precise dimensions of the inflation are not easy to establish, but in Spain, for example, prices at the end of the century were between three and four times greater than those at the beginning; elsewhere in France, England, Holland, Alsace, Italy, Germany, and Poland, they ranged from two to three times their 1500 level at least. The movement was not uniform or smooth. It began about 1480 in Germany, England, and Poland, about 1500 in Spain, and about 1530 in Italy (which generally experienced a smaller rise than most other places). Everywhere, the sharpest rise came between about 1540 and the 1570s. The inflation ended, as it began, at different times in different places: in Germany as early as the 1620s and in England and Poland by the 1640s. After the mid-

17th century the price level almost everywhere remained steady or even declined for more than a century.

Basing their arguments on the theory that prices move broadly in relation to the volume of money in circulation and the velocity of that circulation, some observers from the 16th to the 20th century have seen a major cause of inflation in the swelling stream of silver from Spanish America. They include Francisco López de Gomara, a Spanish historiographer, writing in 1558, Jean Bodin in his dispute with Malestroit in 1568–78, the author of the famous English *Dialogue of the Common Weal* (1581), and the followers of Earl Jefferson Hamilton, a U.S. scholar who has done most to reveal the dimensions of the problem in the 20th century.

Imports of silver from Spanish America

The sequence of events was that the conquistadores, finding silver and gold plates and statues relatively common in Spanish America, began very soon after their arrival to mine precious metals. Boom levels were reached by 1550, after the discovery of rich silver fields in Bolivia and Mexico coincided with a new method of extracting silver through the use of mercury. Up to the 1630s large supplies of silver were mined and sent to Europe. The peak years, about the turn of the century, saw an annual flow of about 36,000,000 pesos. Thereafter, the flow began to dwindle until it fell to a relative trickle by 1660. During this whole period some 18,000 tons of silver (roughly the equivalent of one year's output from present-day South African mines) was added to Europe's monetary stocks. Of the total, about one-fifth was taken by the Spanish crown, some circulated in Spanish America, while the rest was convoyed twice a year to Seville, from where it was distributed throughout Spain and ultimately throughout Europe and to the East.

Causes of inflation. The relationship of this flow to the inflation nevertheless raises certain difficulties. It has been noted, for example, that in some countries the price rise began at least 75 years before 1570, yet Hamilton himself is inclined to minimize the export of bullion from Spain before the latter years of the 16th century. Again, the dimensions of the price rise varied not only between countries but also between commodities. Thus, in general, every price investigation has confirmed that grain and other food prices rose a good deal more than metals, textiles, industrial goods, and colonial commodities, such as spices.

Clearly the influx of silver from the New World does not answer all the questions about the inflation. The growing output of the mines of Germany between 1480 and 1540, which reached its peak in the early 1540s, may help to explain why German and Polish prices rose earlier than Spanish prices, but it is still necessary to explain the impetus behind the continental mining of silver itself; and the demand for silver detracts from the purely monetary and supply side of the problem. Turning to the demand side of the equation, the important increases in population and changes in income distribution obviously played a role. Everywhere in Spain, Germany, Poland, Holland, and England, the larger cities show a rapid growth. Recent research has tended to suggest that the most important nonmonetary change that made it possible for increased supplies of bullion to infiltrate the economy and drive up prices was a general increase of population, associated with growing demand, larger capital investment, and larger national incomes.

Effects of the price inflation. All over Europe landlords' profits grew and with them rural investment and consumption. Customary tenants and freeholders also frequently enjoyed the benefits of rising prices. Entrepreneurs, businessmen, and speculators reaped some harvest from the gap that often opened between prices and costs, though wages sooner or later caught up with prices. Everywhere governments had to face increased expenditure, and, since their incomes were slow to grow, they often added to the inflationary pressures by debasing their coinage, as did the English Tudors in the 1540s and 1550s. Such debasements reached their climax in the notorious monetary manipulations known as the *Kipper und Wipper Zeit* (see-saw time) of the early 17th century. Larger rural incomes together with larger urban incomes promoted the production and consumption of more luxury goods. Thus, any satisfactory explanation of the inflation must include these demand factors as well as the changes in monetary supply.

Debasement of the coinage

Few contemporaries grasped the underlying causes of the inflation. Naturally, they tended to blame it all on human wickedness and greed, passing legislation to hold down prices, rents, and wages, all to little effect. The inflation and its social consequences continued. Generally, those who could adjust their incomes to the changing price level benefitted most, such as landlords who could force or persuade their tenants to accept flexible commercial leases or tenancies; tenants with security of tenure, who could enjoy higher profits if they had surplus produce to sell, while still paying fixed rents; and tradesmen, who could put up prices and keep costs down. The sufferers were landowners, whose *rentier* incomes were limited by custom, and wage earners whose wages stuck at customary levels while the price of necessities rose. It must be remembered, however, that the wage earner completely reliant on a money wage was still the exception rather than the rule in most societies. Much payment in kind was still in force. Many workers were still rural peasants with a small holding to provide them with at least a subsistence living.

The decline of Spain. It has been argued that the decline of Spain in the 17th century from its 16th-century peak of power and prosperity was fundamentally due to economic causes. The silver influx is claimed to have brought about a high degree of inflation, and, since costs rose with prices, Spain suffered a loss of economic strength that brought about political decline. This theory is not universally accepted, and other historians have argued that Spain's problems arose from the misguided determination of Philip II and his successors to burden Spain with such heavy costs of war and extravagance that the consequential high taxation and poverty drove out all enterprise and large numbers of valuable skilled workers until Spain was to become a proud but arid desert, inhabited by bankrupts, monks, and paupers. This picture, though in some respects exaggerated, gains some credibility from comparisons with other countries, such as the Dutch republic; for there is a fair measure of agreement that the same price inflation that is alleged to have ruined Spain was, at the very same time, exerting a powerful natural influence that added daily to the wealth of the Dutch. (C.H.Wi.)

THE RENAISSANCE

The Italian Renaissance. *Development of the Italian cities.* Medieval Italy was a land of cities. The urban imprint of Roman times, never totally erased during some 500 years of barbarian invasions and settlement, began to reassert itself in Italy by the 10th century. New towns and old ones newly revived began to dot the spiny Italian landscape—striking creations of a population that was burgeoning in numbers and brimming with new energies. As in Roman times, the medieval Italian town lived in close relation to its surrounding rural area, or *contado;* Italian city folk seldom relinquished their ties to the land from which they and their families had sprung. Rare was a successful tradesman or banker who not invest some of his profits in his family farm or a rural noble who did not spend part of his year in his tower house inside city walls. In Italian towns, nobles, merchants, *rentiers,* and skilled craftsmen lived and worked side by side, fought in the same militia, and married into each other's families. Social hierarchy there was, but it was a tangled system with no simple division between noble and commoner, between landed and commercial wealth. That nobles took part in civic affairs helps explain the early militancy of the townfolk in resisting the local bishop, who was usually the principal claimant to power in the community. Political action against a common enemy tended to fuse townspeople with a sense of community and civic loyalty. By the end of the 11th century civic patriotism began to express itself in literature; city chronicles combined fact and legend to stress a city's Roman origins and, in some cases, its inheritance of Rome's special mission to rule. Such motifs reflect the cities' achievement of autonomy from their respective episcopal or secular feudal overlords and, probably, the growth of rivalries between neighbouring communities.

Urban growth

This in turn was part of the expansion into the neighbouring countryside, with the smaller and weaker towns submitting to the domination of the larger and stronger. As the activity of the towns became more complex, sporadic political action was replaced by permanent civic institutions. Typically, the first of these was an executive magistracy, named the consulate (to stress the continuity with republican Rome). In the late 11th and early 12th centuries, this process—consisting of the establishment of juridical autonomy, the emergence of a a permanent officialdom, and the spread of power beyond the walls of the city to the *contado* and neighbouring towns—was completed for about a dozen Italian centres and underway for some dozens more; the loose urban community was becoming a corporate entity, or *commune;* the city was becoming a city-state. The typical 13th century city-state was a republic administering a territory of dependent towns; whether it was a democracy is a question of definition. The idea of popular sovereignty existed in political thought and was reflected in the practice of calling a *parlamento,* or mass meeting, of the populace in times of emergency; but in none of the republics were the people as a whole admitted to regular participation in government. On the other hand, the 13th century saw the establishment, after considerable struggle, of assemblies in which some portion of the citizenry, determined by property and other qualifications, took part in debate, legislation, and the selection of officials. Most offices were filled by citizens serving on a rotating, short-term basis. If the almost universal obligation of service in the civic militia is also considered, it becomes clear that participation in the public life of the commune was shared by a considerable part of the population. Moreover, most of the city republics were small enough (in 1300 Florence, one of the very largest, had perhaps 100,000 people; Padua, nearer the average, had about 15,000) so that public business was conducted by and for citizens who knew each other, and civic issues were a matter of widespread and intense personal concern.

A darker but ever-present side of this intense community involvement was conflict. It became a cliché of contemporary observers that when the citizens were not fighting wars with their neighbours they were fighting each other. Machiavelli explained this as the result of the natural enmity between nobles and "the people—the former desiring to command, the latter unwilling to obey." Although this is too simple, it contains an essential truth: the basic problem was the unequal distribution of power and privilege, complicated by the persistence of violent feudal ways (the nobles' militant style was widely imitated as a standard of behaviour). The inability to contain conflict within peaceable limits resulted in bloody strife between members of rival factions, such as the Guelfs and Ghibellines, with those on the losing side often forced into exile and suffering the confiscation of their property.

Enmity between nobles and the people

During the 14th century a number of cities, despairing of finding a solution to the problem of civic strife, were turning from republicanism to *signoria,* the rule of one man. The *signore,* or lord, was usually a member of a local feudal family that was also a power in the commune; thus, lordship did not appear to be an abnormal development, particularly if the *signore* chose, as most did, to rule through existing republican institutions. Sometimes a *signoria* was established as the result of one noble faction's victory over another, while in a few cases a feudal noble who had been hired by the republic as its *condottiere,* or military captain, became its master. Whatever the process, hereditary lordship had become the common condition and free republicanism the exception by the late 14th century. Contrary to what Burckhardt believed, Italy in the 14th century had not shaken off feudalism. In the south, feudalism was entrenched in the loosely centralized Kingdom of Naples, successor state to the Hohenstaufen and Norman kingdoms. In central and northern Italy, feudal lordship and knightly values merged with medieval communal institutions to produce the typical state of the Renaissance—a state that was a compromise between conflicting tendencies. Where the nobles were excluded by law from political participation in the commune, as in the Tuscan cities of Florence, Siena, Pisa, and Lucca, parliamentary republicanism had a longer life; but even these bastions of liberty had their intervals of disguised or open lordship. The great maritime republic of Venice reversed the usual process by increasing the powers of its councils at the expense of the doge (Latin *dux,* leader). But Venice never had a feudal nobility, only a merchant aristocracy that called itself noble and jealously guarded its hereditary sovereignty against incursions from below.

Wars of expansion. There were new as well as traditional elements in the Renaissance city-state. Changes in the political and economic situation affected the evolution of government, while the growth of the Humanist movement influenced conceptions of citizenship, patriotism, and civic history. The decline in the ability of both the empire and the papacy to dominate Italian affairs as they had done in the past left each state free to pursue its own goals within the limits of its resources. These goals were, invariably, the security and power of each state vis-à-vis its neighbours. Diplomacy became a skilled game of experts; rivalries were deadly, and warfare was endemic. Because the costs of war were all-consuming, particularly as mercenary troops replaced citizen militias, the states had to find new sources of revenue and to develop methods of securing public credit. New officials were required to administer these revenues and to gather information and keep records. The city-state came to take over many of the functions formerly performed by associations of private citizens—the kinship groups, tower consortiums, guilds, and political parties that had regulated social relations—leaving individuals to confront the state alone and without intermediaries. If Renaissance man became conscious of himself as an individual, as Burckhardt declared, he also became inescapably conscious of his relation to the state, which became father, mother, and family to everyone under its jurisdiction.

In place of the prevailing anarchy, the Italian states had begun to evolve a new pattern by the 14th century. In place of the dual orbits of empire and papacy in which most of them had revolved, regional powers emerged, with the stronger or more ambitious from time to time bidding for domination of the peninsula. The most sustained effort of aggrandizement was that of Milan under the lordship of the Visconti. In the 1380s and 1390s Gian Galeazzo Visconti pushed Milanese hegemony eastward as far as Padua, at the very doorstep of Venice, and southward to the Tuscan cities of Lucca, Pisa, and Siena and even to Perugia in papal territory. Some believed that Gian Galeazzo meant to be king of Italy; whether or not this is true, he would probably have overrun Florence, the last outpost of resistance in central Italy, had he not died suddenly in 1402 leaving a divided inheritance and much confusion. In the 1420s, under Filippo Maria, Milan began to expand again; but by then Venice, with territorial ambitions of its own, had joined with Florence to block Milan's advance, while the other Italian states took sides or remained neutral according to their own interests. The mid-15th century saw the Italian peninsula embroiled in a turmoil of intrigues, plots, revolts, wars, and shifting alliances, of which the most sensational was the reversal that brought the two old enemies, Florence and Milan, together against Venetian expansion. This "diplomatic revolution," engineered by Cosimo de' Medici (1389–1464), the unofficial head of the Florentine republic, is the most significant illustration of the emergence of balance of power diplomacy in Renaissance Italy.

Emergence of regional powers

Italian Humanism and scholarship. As seen above, the notion that ancient wisdom and eloquence lay slumbering in the Dark Ages until awakened in the Renaissance was the creation of the Renaissance itself. The idea of the revival of classical antiquity was one of those great myths, comparable to the idea of the universal civilizing mission of imperial Rome or to the idea of progress in a modern industrial society, by which an era defines itself in history. Like all such myths, it was a blend of fact and invention. Classical thought and style permeated medieval culture in ways past counting. Most of the authors known to the Renaissance were known to the Middle Ages as well, while the classical works "discovered" by the Humanists were not originals but medieval copies preserved in monastic

Earlier revivals of classical antiquity

or cathedral libraries. Moreover, the Middle Ages had produced at least two earlier revivals of classical antiquity. The so-called Carolingian Renaissance of the late 8th and 9th centuries saved many ancient works from destruction or oblivion, passing them down to posterity in its beautiful minuscule script (which influenced the Humanist scripts of the Renaissance). A 12th-century Renaissance saw the revival of Roman law, Latin poetry, and Greek science, including almost the whole corpus of Aristotelian writings known today.

Growth of literacy. Nevertheless, the classical revival of the Italian Renaissance was so different from these earlier movements in spirit and substance that the Humanists understandably felt it was original and unique. During most of the Middle Ages, classical studies and virtually all intellectual activities were carried on by churchmen, usually members of the regular orders. In the Italian cities this monopoly was partially breached by the growth of a literate laity with some taste and need for literary culture. New professions reflected the growth of both literary and specialized lay education—the *dictatores,* or teachers of practical rhetoric, and lawyers, and the ever present notary (a combination of accountant, solicitor, and public recorder). These, and not Burckhardt's wandering scholar-clerics, were the true predecessors of the Humanists.

In Padua a kind of early Humanism emerged, flourished, and declined between the late 13th and early 14th centuries. Paduan classicism was a product of the vigorous republican life of the commune, and its decline coincided with the loss of the city's liberty. A group of Paduan jurists, lawyers, and notaries—all trained as *dictatores*—developed a taste for classical literature that probably stemmed from their professional interest in Roman law and their affinity for the history of the Roman republic. The most famous of these Paduan classicists was Albertino Mussato, a poet, historian, and playwright, as well as lawyer and politician, whose play *Ecerinis,* modelled on Seneca, has been called the first Renaissance tragedy. By reviving several types of ancient literary forms and by promoting the use of classical models for poetry and rhetoric, the Paduan Humanists helped make the 14th-century Italians more conscious of their classical heritage; in other respects, however, they remained close to their medieval antecedents, showing little comprehension of the vast cultural and historical gulf that separated them from the ancients.

Language and eloquence. It was Petrarch who first understood fully that antiquity was a civilization apart and, understanding it, outlined a program of classically oriented studies that would lay bare its spirit. The focus of Petrarch's insight was language: if classical antiquity was to be understood in its own terms it would be through the speech with which the ancients had communicated their thoughts. This meant that the languages of antiquity had to be studied as the ancients had used them and not as vehicles for carrying modern thoughts. Thus, grammar, which included the reading and careful imitation of ancient authors from a linguistic point of view, was the basis of Petrarch's entire program.

Eloquence is the union of elegance, power, and virtue

From the mastery of language one moved on to the attainment of eloquence. For Petrarch, as for Cicero, eloquence was not merely the possession of an elegant style, nor yet the power of persuasion, but the union of elegance and power together with virtue. One who studied language and rhetoric in the tradition of the great orators of antiquity did so for a moral purpose—to persuade men to the good life—for, said Petrarch in a dictum that could stand as the slogan of Renaissance Humanism, "it is better to will the good than to know the truth."

The humanities. To will the good, one must first know it; and so there could be no true eloquence without wisdom. According to Leonardo Bruni, a leading Humanist of the next generation, Petrarch "opened the way for us to show in what manner we might acquire learning." Petrarch's union of rhetoric and philosophy, modelled on the classical ideal of eloquence, provided the Humanists with an intellectual dignity and a moral ethos lacking to the medieval *dictatores* and classicists. It also pointed the way toward a program of studies—the *studia humanitatis*—by which the ideal might be achieved. As elaborated by Bruni, Pier Paolo Vergerio, and others, the notion of the humanities was based on classical models—the tradition of a liberal arts curriculum conceived by the Greeks and elaborated by Cicero and Quintilian. Medieval scholars had been fascinated by the notion that there were seven liberal arts, no more and no less, although they did not always agree as to which they were. The Humanists had their own favourites, which invariably included grammar, rhetoric, poetry, moral philosophy, and history, with a nod or two toward music and mathematics. They also had their own ideas about methods of teaching and study. They insisted upon the mastery of Classical Latin and, where possible, Greek, which began to be studied again in the West in 1397, when the Greek scholar Manuel Chrysoloras was invited to lecture in Florence. They also insisted upon the study of classical authors at first hand, banishing the medieval textbooks and compendiums from their schools. This greatly increased the demand for classical texts, which was first met by copying manuscript books in the newly developed Humanistic scripts and then, after the mid-15th century, by the method of printing with movable type. Thus, while it is true that most of the ancient authors were already known in the Middle Ages, there was an all-important difference between circulating a book in many copies to a reading public and jealously guarding a single exemplar as a prized possession in some remote monastery library.

The term humanist (Italian *umanista,* Latin *humanista*) first occurs in 15th-century documents to refer to a teacher of the humanities. Humanists taught in a variety of ways. Some founded their own schools, as Vittorino da Feltre did in Mantua in 1423 and Guarino Veronese in Ferrara in 1429, where students could study the new curriculum at both elementary and advanced levels. Some Humanists taught in universities, which, while remaining strongholds of specialization in law, medicine, and theology, had begun to make a place for the new disciplines by the late 14th century. Still others were employed in private households, as was the great Politian (Angelo Poliziano), who was tutor to the Medici children as well as a university professor.

Literary works of the Humanists

Formal education was only one of several ways in which the Humanists shaped the minds of their age. Many were themselves fine literary artists who exemplified the eloquence they were trying to foster in their students. Renaissance Latin poetry, for example, nowadays dismissed—usually unread—as imitative and formalistic, contains much graceful and lyrical expression by such Humanists as Politian, Giovanni Pontano, and Jacopo Sannazzaro. In drama, Politian, Pontano, and Pietro Bembo were important innovators, and the Humanists were in their element in the composition of elegant letters, dialogues, and discourses. By the late 15th century, Humanists were beginning to apply their ideas about language and literature to composition in Italian as well as in Latin, demonstrating that the "vulgar" tongue could be as supple and as elegant in poetry and prose as was Classical Latin.

Classical scholarship. Not every Humanist was a poet, but most were classical scholars. Classical scholarship consisted of a set of related, specialized techniques by which the cultural heritage of antiquity was made available for convenient use. Essentially, in addition to searching out and authenticating ancient authors and works, this meant editing—comparing variant manuscripts of a work, correcting faulty or doubtful passages, and commenting in notes or in separate treatises on the style, meaning, and context of an author's thought. Obviously, this demanded not only superb mastery of the languages involved and a command of classical literature but also a knowledge of the culture that formed the ancient author's mind and influenced his writing. Consequently, the Humanists created a vast scholarly literature devoted to these matters and instructive in the critical techniques of classical philology, the study of ancient texts.

The study of man through his behaviour

The Humanists' conception of man. Beginning with Petrarch, Humanist thought approached the subject of man through introspection and an examination of actual behaviour rather than through doctrinaire formulas. Striving for the good life, yet torn by selfish passions, caught

between desires for immortality and for earthly fame, alternating between glory and despair—this was man as Petrarch knew him because the subject of his inquiry was himself. As orators, poets, historians, and moralists the Humanists dealt with man acting, willing, creating his own beauty and his own worldly destiny. Humanist psychology tended to stress the volitional and emotional rather than the rational side of man's nature. Reason was important but limited; it could neither master the passions nor grasp the ultimate mysteries. Reason was subordinate to the will, and the will was free to determine man's destiny. To quote Petrarch again, "It is better to will the good than to know the truth."

This voluntarist, moralist conception of man reflects the Humanists' function as spokesmen of an urban, literate laity. To those who had found rewarding social roles in the public world of the city-state, the conception of human life elaborated by medieval theologians was no more useful than an educational system in the service of monastic and priestly values. The Humanists' conception of man as actor and creator was the counterpart of their definition of a new curriculum of liberal studies, with its emphasis upon the relation between culture and the good life. Here the Renaissance myth of antiquity served the Humanists in two ways: first, as a summation of, or way of access to, an extraordinary range of experience (since they believed that the ancients had achieved everything worthwhile in this life); second, as a legitimating authority for modes of thought and action that were unknown or disapproved of in traditional Christian doctrine.

The Humanists were secularists who found in classical antiquity a conception of human life that was congenial and supportive; but they were also Christians, and in their use of ancient thought they showed an independent judgment that belies the old notion of a "pagan Renaissance." In fact, some of their favourite sources were post-classical and Christian. Even the most celebrated of Renaissance themes, the idea of "the dignity of man," best known in the *Oration* of Pico della Mirandola, was derived in part from St. Augustine and other Church Fathers. Created in the image and likeness of God, man was free to shape his own destiny; but the definition of that destiny was very much in the Christian tradition:

> You will have the power to sink to the lower forms of life, which are brutish. You will have the power, through your own judgment, to be reborn into the higher forms, which are divine.

Political thought. Pico envisioned the potential divinity of man; Niccolo Machiavelli looked unblinkingly at his actual behaviour; Francesco Guicciardini saw him as a victim of that behaviour.

The author of *The Prince,* a treatise on how to get power and keep it, Machiavelli so shocked his readers—not because he described behaviour that was unfamiliar to them but because he dared to advocate it—that they coined his name into synonyms for the devil (Old Nick) and for crafty, unscrupulous tactics (Machiavellian). No other name but that of Borgia so invariably evokes the image of the wicked Renaissance, and, indeed, Cesare Borgia was Machiavelli's chief model for *The Prince.* Yet Machiavelli, too, was influenced by Humanism—in particular, by Humanist voluntarist psychology and faith in human freedom. He was also a devotee of the cult of antiquity, which he reverenced chiefly in the ancient historians for their examples of political wisdom and military valour. From the example of Rome he derived the laws of political behaviour by which, he believed, his countrymen must live if they were to recover their civic virtue and free themselves from the yoke of the "barbarians" who were overrunning Italy.

Machiavelli's ideas of the uses of the ancient past were indebted to a Florentine tradition of Civic Humanism that, by his time, was a century old. A line of Humanists had expounded the ideals of republican citizenship and the meaning of classical studies for the public life. In the years of Milanese aggression against Florence, Coluccio Salutati, serving as chancellor of the republic, had rallied the Florentines by reminding them of the legend that their city was "the daughter of Rome" and urging upon them the Roman legacy of justice and liberty. Salutati's pupil, Leonardo Bruni, who also served as chancellor, expounded many of the tenets of Civic Humanism in his *History of the Florentine People* and his panegyrics of Florence. The roots of Florentine liberty, he maintained, were deep in the soil of Tuscany, where, even before the rise of Rome, the Etruscans had founded free cities. In Florence, declared Bruni, equality was recognized in justice and opportunity for all citizens, while the claims of individual excellence were rewarded by preferment to public office and in public honours. Thus, the relation between freedom and achievement was close, and this explained Florence's pre-eminence in culture as well as in political liberty. Florence, observed Bruni, was the home of Italy's greatest poets, the pioneer in both vernacular and Latin literature, and the seat of the Greek revival as well as of eloquence. It was, in short, the centre of the *studia humanitatis,* those studies by and for a free man.

Civic Humanism

As a theory of political life, Civic Humanism represented the ideal rather than the reality of 15th-century communal politics. Even in Florence, where, after 1434, the Medici family took a grip on the city's republican institutions, the emphasis shifted from activism to the kind of Utopian mysticism represented by Pico's *Oration* and to the millennialist fantasies most vividly expressed in the late-15th-century preaching of Fra Girolamo Savonarola. Nevertheless, in providing ways of thinking about the republic and its history, the Humanist chancellors had contributed to a tradition of civic thought that the Medici had failed to extinguish. Machiavelli, himself a secretary in the chancellery of the anti-Medicean republic of 1494–1512 and a member of the group of florentine patricians and Humanists that met in the Rucellai gardens to talk about politics, was deeply influenced by it. In his *Florentine History,* in the *Art of War,* and above all in the *Discourses on the First Ten Books of Livy,* Machiavelli is best seen in the context of the Florentine Humanist tradition. In the *Discourses,* where Machiavelli examines the whole range of political possibilities, it is clear that his own preference in forms of government is for republics. The key to this preference is his idea of man as a political being, a conception he shared with Bruni and indeed with Bruni's source—Aristotle. To participate in the public life, to interact with one's fellowmen in making decisions, was to fulfill one's human nature. For this the best setting was the well-ordered republic, in which no one had a monopoly of power and citizens were devoted to the welfare and service of the community. But well-ordered republics were scarce in 16th-century Italy, as scarce as *virtù,* the political energy that made them possible.

Conflict between Christian and political morality

Machiavelli believed that this scarcity derived from the modern disposition to follow the precepts of conventional Christian morality rather than the more relevant example of ancient political experience. He said that in politics antiquity was more admired than imitated, whereas the teachings of religion had made states weak and sluggish. His solution was to expound Livy, the classical historian of the Roman Republic, in order to determine which examples were applicable to his own time. In this, he saw himself as setting out on a new route, but one that paralleled the paths already taken by modern jurisprudence and medicine, which had already created science by studying ancient laws and the knowledge of the ancient physicians; he proposed to do the same for politics. Thus, for Machiavelli, statecraft was a discipline based on timeless rules or laws and was no more to be subordinated to Christian ethics than were jurisprudence or medicine. The simplest example of the conflict between Christian and political morality is provided by warfare, where the use of deception, so detestable in every other kind of action, is necessary, praiseworthy, and even glorious. Machiavelli's political morality may be summed up in the *Discourses,* where, commenting upon a Roman defeat, he writes,

> This is worth noting by every citizen who is called upon to give counsel to his country, for when the very safety of the country is at stake there should be no question of justice or injustice, of mercy or cruelty, of honour or disgrace, but putting every other consideration aside, that course should be followed which will save her life and liberty.

Machiavelli's own country was Florence; when he wrote that he loved his country more than he loved his soul, he was consciously forsaking Christian ethics for the morality of civic virtue. His friend and countryman Francesco Guicciardini shared his political morality and his concern for politics but lacked his faith that a knowledge of ancient political wisdom would redeem the liberty of Italy. Guicciardini was an upper class Florentine who chose a career in public administration and devoted his leisure to writing history and reflecting on politics. He was steeped in the Humanist traditions of Florence and was a dedicated republican, despite the fact—or perhaps because of it—that he spent his entire career in the service of the Medici and rose to high positions under them. But Guicciardini, more skeptical and aristocratic than Machiavelli, was also half a generation younger, and he was schooled in an age that was already witnessing the decline of Italian autonomy.

In 1527 Florence revolted against the Medici a second time and established a republic. As a confidant of the Medici, Guicciardini was passed over for public office and retired to his estate. One of the fruits of this enforced leisure was the so-called *Cose fiorentine* (*Florentine Affairs*), an unfinished manuscript on Florentine history. While it generally follows the classic form of Humanist civic history, the fragment contains some significant departures from this tradition. No longer is the history of the city treated in isolation; Guicciardini was becoming aware that the political fortunes of Florence were interwoven with those of Italy as a whole and that the French invasion of Italy was a turning point in Italian history. He returned to public life with the restoration of the Medici in 1530 and was involved in the events leading to the tightening of the imperial grip upon Italy, the humbling of the papacy, and the final transformation of the republic of Florence into a hereditary Medici dukedom. Frustrated in his efforts to influence the rulers of Florence, he retired to his villa to write; but instead of taking up the unfinished manuscript on Florentine history, he chose a subject commensurate with his changed perspective on Italian affairs. The result was his *History of Italy*. Though still in the Humanist form and style, it was in substance a fulfillment of the new tendencies already evident in the earlier work—criticism of sources, great attention to detail, avoidance of moral generalizations, shrewd analysis of character and motive.

The *History of Italy* has rightly been called a tragedy by Felix Gilbert, for it demonstrates how, out of stupidity and weakness, men make mistakes that gradually narrow the range of their freedom to choose alternative courses and thus to influence events until, finally, they are trapped in the web of Fortune. This was already far from the world of Machiavelli, not to mention that of the Civic Humanists. Where the Humanists believed that *virtù* could master Fortune, Guicciardini was skeptical about men's ability to learn from the past and pessimistic about their power to influence the course of their own destinies. All that was left, he believed, was to understand. Guicciardini wrote history to show what men were like and to explain how they had reached their present circumstances. Man's dignity, then, consisted not in the exercise of his will to shape his destiny but in the use of his reason to contemplate and perhaps to tolerate his fate. In taking a new, hard look at the human condition, Guicciardini represents the decline of the Humanist view of man.

Guicciardini's pessimism

Arts and letters. As pointed out above, classicism and the literary impulse went hand in hand. From Lovato Lovati and Albertino Mussato to Politian and Pontano, Humanists wrote Latin poetry and drama with considerable grace and power (Politian wrote in Greek as well), while others composed epistles, essays, dialogues, treatises, and histories on classical models. In fact, it is fair to say that the development of an elegant, nonclassical style of prose writing was the major literary achievement of Humanism and that the epistle was its typical literary form. Petrarch's practice of collecting, reordering, and even rewriting his letters—of treating them as works of art—was widely imitated.

For lengthier discussions the Humanist was likely to compose a formal treatise or a dialogue—a classical form that provided the opportunity to combine literary imagination with the discussion of weighty matters. The most famous example of this type is *The Courtier,* published by Baldassare Castiglione in 1528; a graceful discussion of love, courtly manners, and the ideal education for a perfect gentleman, it had enormous influence all over Europe. Castiglione had a Humanist education, but he wrote *The Courtier* in Italian, the language Bembo chose for his dialogue on love, *Gli Asolani* (1505), and Ludovico Ariosto chose for his delightful epic, *Orlando furioso,* completed in 1516. The vernacular was coming of age as a literary medium.

According to some a life-and-death struggle between Latin and Italian began in the 14th century, while the mortal enemies of Italian were the Humanists, who impeded the natural growth of the vernacular after its brilliant beginning with Dante, Petrarch, and Boccaccio. In this view, the choice of Italian by such great 16th-century writers as Castiglione, Ariosto, and Machiavelli represents the final "triumph" of the vernacular and the restoration of contact between Renaissance culture and its native roots. The reality is somewhat less dramatic and more complicated. Most Italian writers regarded Latin as being as much a part of their culture as the vernacular, and most of them wrote in both languages. It should also be remembered that Italy was a land of powerful regional dialect traditions; until the late 13th century, Latin was the only language common to all Italians. By the end of that century, however, Tuscan was emerging as the primary vernacular, and Dante's choice of it for his *Divine Comedy* ensured its pre-eminence. Of lyric poets writing in Tuscan (hereafter called Italian), the greatest was Petrarch. His *canzoni,* or songs, and sonnets in praise of Laura are revealing studies of the effect of love upon the lover; his *Italia mia* is a plea for peace that evokes the beauties of his native land; his religious songs reveal his deep spiritual feeling.

Emergence of Tuscan as the vernacular

Petrarch's friend and admirer Giovanni Boccaccio is best known for his *Decameron;* but he pioneered in adapting classical forms to Italian usage, including the hunting poem, romance, idyll, and pastoral, whereas some of his themes, most notably the story of Troilus and Cressida, were borrowed by other poets, including Chaucer and Tasso.

The scarcity of first-rate Italian poetry throughout most of the 15th century has caused a number of historians to regret the passing of "il buon secolo," the great age of the language, which supposedly came to an end with the ascendancy of Humanist classicism. For every Humanist who disdained the vernacular, however, there was a Leonardo Bruni to maintain its excellence or a Poggio Bracciolini to prove it in his own Italian writings. Indeed, there was an absence of first-rate Latin poets until the late 15th century, which suggests a general lack of poetic creativity in this period and not of Italian poetry alone. It may be that both Italian and Latin poets needed time to absorb and assimilate the various new tendencies of the preceding period. Tuscan was as much a new language for many as was Classical Latin, and there was a variety of literary forms to be mastered.

With Lorenzo de' Medici the period of tutelage came to an end. The Magnificent Lorenzo, virtual ruler of Florence in the late 15th century, was one of the fine poets of his time. His sonnets show that he had felt Petrarch's influence but transformed it with his own genius. His poetry epitomizes the Renaissance ideal of *l'uomo universale,* the many-sided man. Love of nature, love of women, love of life are the principal themes reflecting the experience of the man. The woodland settings and hunting scenes of his poems suggest how he found relief from a busy public life; his love songs to his mistresses and his bawdy carnival ballads show the other face of a devoted father and affectionate husband. The celebration of youth in his most famous poem was etched with the sad realization that time passes swiftly:

Oh, how fair is youth, and yet how fleeting!
Let yourself be joyous if you feel it:
Of tomorrow there is no certainty—

Florence was only one centre of the flowering of the vernacular. Ferrara saw literature and art flourish under the patronage of the ruling Este family and before the end of

the 15th century counted at least one major poet, Matteo Boiardo, author of the *Orlando innamorato,* an epic of Roland. A blending of the Arthurian and Carolingian epic traditions, Boiardo's *Orlando* inspired Ludovico Ariosto to take up the same themes. The result was the finest of all Italian epics, *Orlando furioso.* The ability of the medieval epic and folk traditions to inspire the poets of such sophisticated centres as Florence and Ferrara suggests that Humanist disdain for the Dark Ages notwithstanding, Renaissance Italians did not allow classicism to cut them off from their medieval roots.

The northern Renaissance. *Political, economic, and social background.* In 1494 King Charles VIII of France led an army southward over the Alps, seeking the Neapolitan crown and glory. Many believed that this barely literate gnome of a man, hunched over his horse, was the Second Charlemagne, whose coming had been long predicted by French and Italian prophets. Apparently, Charles himself believed it; it is recorded that when he was chastised by the fiery preacher Savonarola for delaying his divine mission of reform and crusade in Florence, the King burst into tears and soon went on his way. He found the Kingdom of Naples easy to take and impossible to hold; frightened by local uprisings, by a new Italian coalition, and by the massing of Spanish troops in Sicily, he left Naples in the spring of 1495, bound not for the Holy Land, as the prophecies had predicted, but for home, never to return to Italy. In 1498 Savonarola was tortured, hanged, and burned as a false prophet for predicting that Charles would complete his mission. Conceived amid dreams of chivalric glory and crusade, the Italian expedition of Charles VIII was the venture of a medieval king—romantic, poorly planned, and totally irrelevant to the real needs of his subjects.

New phase of European politics

The French invasion of Italy marked the beginning of a new phase of European politics, during which the Valois kings of France and the Habsburgs of Germany fought each other, with the Italian states as their reluctant pawns. For the next 60 years the dream of Italian conquest was pursued by every French king, none of them having learned anything from Charles VIII's misadventure except that the road southward was open and paved with easy victories. For even longer Italy would be the keystone of the arch that the Habsburgs tried to erect across Europe from the Danube to the Strait of Gibraltar in order to link the Spanish and German inheritance of Charles V. In destroying the autonomy of Italian politics, the invasions also ended the Italian state system, which was absorbed into the larger European system that now took shape. Its members adopted the balance of power diplomacy first evolved by the Italians as well as the Italian practice of using resident ambassadors who combined diplomacy with the gathering of intelligence by fair means or foul. In the art of war, also, the Italians were the schoolmasters of Europe, with their innovations in the use of mercenary troops, cannonry, bastioned fortresses, and field fortification. French artillery was already the best in Europe by 1494, whereas the Spaniards developed the *tercio,* an infantry unit that combined the most effective field fortifications and weaponry of the Italians and Swiss.

Thus, old and new ways were fused in the bloody crucible of the Italian Wars. Rulers who lived by medieval codes of chivalry adopted Renaissance techniques of diplomacy and warfare to satisfy their lust for glory and dynastic power. Even the lure of Italy was an old obsession; only the size and vigour of the 16th-century expeditions were new. Rulers were now able to command vast quantities of men and resources because they were becoming masters of their own domains. The nature and degree of this mastery varied according to local circumstances; but all over Europe the New Monarchs, as they are called, were reasserting kingship as the dominant form of political leadership after a long period of floundering and uncertainty.

By the end of the 15th century the Valois kings of France had expelled the English from all their soil except for the port of Calais, concluding the Hundred Years' War (1453); had incorporated the fertile lands of the duchy of Burgundy to the east and of Brittany to the north; and had extended the French kingdom from the Atlantic and the English Channel to the Pyrenees and the Rhine. To rule this vast territory they created a professional machinery of state, converting wartime taxing privileges into permanent prerogative, freeing their royal council from supervision by the States General, appointing a host of officials who crisscrossed the kingdom in the service of the crown, and establishing their right to appoint and tax the French clergy. They did not achieve anything like complete centralization; but in 1576 Jean Bodin was able to write, in his *Six Books of the Commonweal,* that the king of France had absolute sovereignty because he alone in the kingdom had the power to give law unto all of his subjects in general and to every one of them in particular.

Absolute sovereignty

Bodin might also have made his case by citing the example of another impressive autocrat of his time, Philip II of Spain. Though descended from warrior kings, Philip spent his days at his writing desk poring over dispatches from his governors in the Low Countries, Sicily, Naples, Milan, Peru, Mexico, and the Philippines and drafting his orders to them in letters signed "I the King." The founding of this mighty empire went back more than a century to 1469, when Ferdinand II of Aragon and Isabella of Castile brought two great Hispanic kingdoms together under a single dynasty. Castile, an arid land of sheepherders, great landowning churchmen, and crusading knights, and Aragon, with its Catalan miners and its strong ties to Mediterranean Europe, made uneasy partners; but a series of rapid and energetic actions forced the process of national consolidation and catapulted the new nation into a position of world prominence for which it was poorly prepared. Within the last decade of the 15th century the Spaniards took the kingdom of Navarre in the north; stormed the last Muslim stronghold in Spain, the kingdom of Granada; and launched a campaign of religious unification by pressing tens of thousands of Muslims and Jews to choose between Baptism and expulsion, at the same time establishing a new Inquisition under royal control. They also sent Columbus on voyages of discovery to the Western Hemisphere, thereby opening a new frontier just as the domestic frontier of Reconquest was closing. Finally, the crown linked its destinies with the Habsburgs by a double marriage, thus projecting Spain into the heart of European politics. In the following decades Castilian *hidalgos* (lower nobles), whose fathers had crusaded against the Moors in Spain, streamed across the Atlantic to make their fortunes out of the land and sweat of the American Indians, while others marched in the armies and sailed in the ships of their king, Charles I, who, as Charles V, was elected Holy Roman emperor in 1519 at the age of 19. In this youth the vast dual inheritance of the Spanish and Habsburg empires came together. The grandson of Ferdinand and Isabella on his mother's side and of the emperor Maximilian I on his father's, Charles was duke of Burgundy, head of five Austrian dukedoms (which he ceded to his brother), king of Naples, Sicily, and Sardinia, and claimant to the duchy of Milan as well as king of Aragon and Castile and German king and emperor. To administer this enormous legacy, he presided over an ever-increasing bureaucracy of viceroys, governors, judges, military captains, and an army of clerks. The New World lands were governed by a separate Council of the Indies after 1524, which, like Charles' other royal councils, combined judicial, legislative, military, and fiscal functions.

The empire under Charles V

The yield in American treasure was enormous, especially after the opening of the silver mines of Mexico and what is now Bolivia halfway through the 16th century. The crown skimmed off a lion's share—usually a fifth—which it paid out immediately to its creditors because everything Charles could raise by taxing or borrowing was sucked up by his wars against the French in Italy and Burgundy, the Protestant princes in Germany, the Turks on the Austrian border, and the Barbary pirates in the Mediterranean. By 1555 both Charles and his credit were exhausted, and he began to relinquish his titles—Spain and the Netherlands to his son Philip, Germany and the imperial title to his brother Ferdinand I. American silver did little for Spain except to pay the wages of soldiers and sailors; the goods and services that kept the Spanish armies in the field and the ships afloat were largely supplied by foreigners,

who reaped the profits. But for the rest of the century Spain continued to dazzle the world, and few could see the chinks in the armour; for this was an age of kings, in which bold deeds, not balance sheets, made history.

The growth of centralized monarchy claiming absolute sovereignty over its subjects may be observed in other places, from the England of Henry VIII on the extreme west of Europe to the Muscovite kingdom of Ivan III the Great on its eastern edge, for the New Monarchy was one aspect of a more general phenomenon—a great recovery of energy that surged through Europe in the 15th century. No single cause can be adduced to explain it. Some historians believe it was simply the upturn in the natural cycle of growth: the great medieval population boom had overextended Europe's productive capacities; the depression of the 14th and early 15th centuries had corrected this condition through famines and epidemics, leading to depopulation; now the cycle of growth was beginning again.

The effects of increased food production and new technology

Once more, growing numbers of people, burgeoning cities, and ambitious governments were demanding food, goods, and services—a demand that was met by both old and new methods of production. In agriculture, the shift toward commercial crops such as wool and grains, the investment of capital, and the emancipation of servile labour completed the transformation of the manorial system already in decline. (In eastern Europe, however, the formerly free peasantry was now forced into serfdom by an alliance between the monarchy and the landed gentry, as huge agrarian estates were formed to raise grain for an expanding Western market.) Manufacturing boomed, especially of those goods used in the outfitting of armies and fleets—cloth, armour, weapons, and ships. New mining and metalworking technology made possible the profitable exploitation of the rich iron, copper, gold, and silver deposits of central Germany, Hungary, and Austria, affording the opportunity for large-scale investment of capital.

One index of Europe's recovery is the spectacular growth of certain cities. Antwerp, for example, more than doubled its population in the second half of the 15th century and doubled it again by 1560. Under Habsburg patronage Antwerp became the chief European entrepôt for English cloth, the hub of an international banking network and the principal Western market for German copper and silver, Portuguese spices, and Italian alum. By 1500 the Antwerp Bourse was the central money market for much of Europe. Other cities profited from their special circumstances, too: Lisbon as the home port for the Portuguese maritime empire; Seville, the Spaniards' gateway to the New World; London, the capital of the Tudors and gathering point for England's clothmaking and banking activity; Lyons, favoured by the French kings as a market centre and capital of the silk industry; Augsburg, the principal north-south trade route in Germany and the home city of the Fugger merchant-bankers.

Northern Humanists. Cities were also markets for culture. The resumption of urban growth in the second half of the 15th century coincided with the diffusion of Renaissance ideas and educational values. Humanism offered linguistic and rhetorical skills that were becoming indispensable for nobles and commoners seeking careers in diplomacy and government administration, while the Renaissance ideal of the perfect gentleman was a cultural style that had great appeal in this age of growing courtly refinement. At first those who wanted a Humanist education had to go to Italy, and many foreign names appear on the rosters of the Italian universities. By the end of the century, however, such northern cities as London, Paris, Antwerp, and Augsburg were becoming Humanist centres in their own right. The development of printing, by making books cheaper and more plentiful, also quickened the diffusion of Humanism.

A textbook convention, heavily armoured against truth by constant reiteration, states that northern Humanism—*i.e.,* Humanism outside Italy—was essentially Christian in spirit and purpose, in contrast to the essentially secular nature of Italian Humanism. In fact, however, the program of Christian Humanism had been laid out by Italian Humanists of the stamp of Lorenzo Valla, one of the founders of classical philology, who showed how the critical methods used to study the classics ought to be applied to problems of biblical exegesis and translation as well as church history. That this program only began to be carried out in the 16th century, particularly in the countries of northern Europe (and Spain), is a matter of chronology rather than of geography. In the 15th century the necessary skills, particularly the knowledge of Greek, were possessed by a few scholars; a century later, Greek was a regular part of the Humanist curriculum, and Hebrew was becoming much better known, particularly after Johannes Reuchlin published his Hebrew grammar in 1506. Here, too, printing was a crucial factor, for it made available a host of lexicographical and grammatical handbooks and allowed the establishment of normative biblical texts and the comparison of different versions of the Bible.

Christian Humanism

Christian Humanism was more than a program of scholarship, however; it was fundamentally a conception of the Christian life that was grounded in the rhetorical, historical, and ethical orientation of Humanism itself. That it came to the fore in the early 16th century was the result of a variety of factors, including the spiritual stresses of rapid social change and the inability of the ecclesiastical establishment to cope with the religious needs of an increasingly literate and self-confident laity. By restoring the gospel to the centre of Christian piety, the Humanists believed they were better serving the needs of ordinary people. They attacked Scholastic theology as an arid intellectualization of simple faith, and they deplored the tendency of religion to become a ritual practiced vicariously through a priest. They also despised the whole late-medieval apparatus of relic mongering, hagiology, indulgences, and image worship, and they ridiculed it in their writings, sometimes with devastating effect. According to the Christian Humanists, the fundamental law of Christianity was the law of love as revealed by Jesus Christ in the Gospel. Love, peace, and simplicity should be the aims of the good Christian and the life of Christ his perfect model. The chief spokesman for this point of view was Desiderius Erasmus, the most influential Humanist of his day. Erasmus and his colleagues were uninterested in dogmatic differences and were early champions of religious toleration. In this they were not in tune with the changing times, for the outbreak of the Reformation polarized European society along confessional lines, with the paradoxical result that the Christian Humanists, who had done so much to lay the groundwork for religious reform, ended by being suspect on both sides—by the Catholics as subversives who (as it was said of Erasmus) had "laid the egg that Luther hatched," and by the Protestants as hypocrites who had abandoned the cause of reformation out of cowardice or ambition. Toleration belonged to the future, after the killing in the name of Christ sickened and passions had cooled.

Christian mystics. The quickening of the religious impulse that gave rise to Christian Humanism was also manifested in a variety of forms of religious devotion among the laity, including mysticism. In the 14th century a wave of mystical ardour seemed to course down the valley of the Rhine, enveloping men and women in the rapture of intense, direct experience of the divine Spirit. It centred in the houses of the Dominican order, where friars and sisters practiced the mystical way of their great teacher, Meister Eckehart. This wave of Rhenish mysticism radiated beyond convent walls to the marketplaces and hearths of the laity. Eckehart had the gift of making his abstruse doctrines understandable to a wider public than was usual for mystics; moreover, he was fortunate in having some disciples of a genius almost equal to his own—the great preacher of practical piety, Johann Tauler, and Heinrich Suso, whose devotional books, such as *The Little Book of Truth* and *The Little Book of Eternal Wisdom,* reached eager lay readers hungry for spiritual consolation and religious excitement. Some found it by joining the Dominicans; others, remaining in the everyday world, joined with like-spirited brothers and sisters in groups known collectively as the Friends of God, where they practiced methodical contemplation, or, as it was widely known, mental prayer. Probably few reached, or even hoped to reach, the ecstasy of mystical union, which was limited to those with the appropriate psychological or spiritual gifts.

Spread of mysticism

Out of these circles came the anonymous *German Theology,* from which Luther was to say that he had learned more about man and God than from any book except the Bible and the writings of St. Augustine.

In the Netherlands the mystical impulse awakened chiefly under the stimulus of another great teacher, Gerhard Groote. Not a monk nor even a priest, Groote gave the mystical movement a different direction by teaching that true spiritual communion must be combined with moral action, for this was the whole lesson of the Gospel. At his death a group of followers formed the Brethren of the Common Life. These were laymen and laywomen, married and single, earning their livings in the world but united by a simple rule that required them to pool their earnings and devote themselves to spiritual works, teaching, and charity. Houses of Brothers and Sisters of the Common Life spread through the cities and towns of the Netherlands and Germany, and a monastic counterpart was founded in the order of Canons Regular of St. Augustine, known as the Windesheim Congregation, which, in the second half of the 15th century numbered some 82 priories. The Brethren were particularly successful as schoolmasters, combining some of the new linguistic methods of the Humanists with a strong emphasis upon Bible study. Among the generations of children who absorbed the New Piety (*devotio moderna*) in their schools were Erasmus, and, briefly, Luther. In the ambience of the *devotio moderna* appeared one of the most influential books of piety ever written, *The Imitation of Christ,* attributed to Thomas à Kempis, a monk of the Windesheim Congregation.

One man whose life was changed by *The Imitation* was the 16th-century Spaniard Ignatius Loyola. After reading it Loyola founded the Society of Jesus, and wrote his own book of methodical prayer, *Spiritual Exercises.* Thus, Spanish piety was in some ways connected with that of the Netherlands; but the extraordinary outburst of mystical and contemplative activity in 16th-century Spain was mainly an expression of the intense religious exaltation of the Spanish people themselves as they confronted the tasks of reform, Counter-Reformation, and world leadership. Spanish mysticism belies the usual picture of the mystic as a withdrawn contemplative, with his or her head literally in the clouds. Not only Loyola but also St. Teresa of Avila and her disciple, St. John of the Cross, were tough, activist Reformers who regarded their mystical experiences as means of fortifying themselves for their practical tasks. They were also prolific writers who could communicate their experiences and analyze them for the benefit of others. This is especially true of St. John of the Cross, whose mystical poetry is one of the glories of Spanish literature.

The growth of vernacular literature. In literature, medieval forms continued to dominate the artistic imagination throughout the 15th century. Besides the vast devotional literature of the period—the books on *The Art of Dying Well,* the saints' lives, and manuals of methodical prayer and spiritual consolation—the most popular reading of noble and burgher alike was the 13th-century love allegory, the *Roman de le rose.* Despite a promising start in the late Middle Ages, literary creativity suffered from the domination of Latin as the language of "serious" expression, with the result that if the vernacular attracted writers, they tended to overload it with Latinisms and artificially applied rhetorical forms. This was the case with the so-called *grande rhetoriqueurs* of Burgundy and France. One exception is 14th-century England, where a national literature made a brilliant showing in the works of William Langland, John Gower, and, above all, Geoffrey Chaucer; but the troubled 15th century produced only feeble imitations. Another exception is the vigorous tradition of chronicle writing in French, distinguished by such eminently readable works as the chronicle of Jean Froissart and the memoirs of Philippe de Commynes. In France, too, around the middle of the 15th century there lived the vagabond François Villon, a great poet about whom next to nothing is known. In Germany *The Ship of Fools,* by Sebastian Brant, was a lone masterpiece.

The 16th century saw a true renaissance of national literatures. In Protestant countries the Reformation had an enormous impact upon the quantity and quality of literary output. If Luther's rebellion destroyed the chances of unifying the nation politically, his translation of the Bible into German created a national language. Biblical translations, vernacular liturgies, hymns, and sacred drama had analogous effects elsewhere. For Catholics, especially in Spain, the Reformation was a time of deep religious emotion expressed in art and literature. On all sides of the religious controversy, chroniclers and historians writing in the vernacular were recording their versions for posterity. Renaissance of national literatures

While the Reformation was providing a subject matter, the Italian Renaissance was providing literary methods and models. The Petrarchan sonnet inspired French, English, and Spanish poets, while the Renaissance Neoclassical drama finally began to end the reign of the medieval mystery play. Ultimately, of course, the works of real genius were the result of a crossing of native traditions and new forms. The Frenchman François Rabelais assimilated all the themes of his day—and mocked them all—in his story of the giants Gargantua and Pantagruel. The Spaniard Miguel de Cervantes, in *Don Quixote,* drew a composite portrait of his countrymen, which caught their exact mixture of idealism and realism. In England, Christopher Marlowe and William Shakespeare used Renaissance drama to probe the deeper levels of their countrymen's character and experiences.

Renaissance science and technology. To the medieval mind, matter was composed of four elements—earth, air, fire, and water—whose combinations and permutations made up the world of visible objects. The cosmos was a series of concentric spheres in motion, the farther ones carrying the stars around in their daily courses; and at the centre of all was the globe of Earth, heavy and static. Motion was either perfectly circular, as in the heavens, or irregular and naturally downward, as on the Earth. The Earth was made up of three landmasses—Europe, Asia, and Africa—and was unknown and uninhabitable in its southern zones. Man, the object of all creation, was composed of four humours—black and yellow bile, phlegm, and blood—and his personality and health were determined by the relative proportions he had of each. The cosmos was alive with a common consciousness, and the stars influenced the course of events as well as the fortunes of men (although the church frowned on this denial of free will). Man might influence the spirits in nature through magic—black for demons, white for the benevolent spirits—although the church preferred that the Christian seek his well-being through faith, the sacraments, and the intercession of Mary and the saints.

These views were an amalgam of Aristotelian physics, Galenic medicine, Ptolemaic astronomy, and Christian theology. Together they ruled man's understanding and directed his experience of phenomena, until they all were overthrown and replaced by the new mechanistic conceptions of Copernicus, William Harvey, Galileo, and Isaac Newton. Only the first of these great scientists was born in the period discussed here as the Renaissance; the Scientific Revolution was largely the achievement of the 17th century. The persistence of the medieval model of the cosmos through the Renaissance ought not, however, to obscure the period's real contributions. Humanist scholarship provided both originals and translations of ancient Greek scientific works—which enormously increased the fund of knowledge in physics, astronomy, medicine, botany, and other disciplines—and presented as well alternative theories to those of Ptolemy and Aristotle. Thus, the revival of ancient science brought heliocentric astronomy to the fore again after almost two millennia. Renaissance philosophers, most notably Jacopo Zabarella, analyzed and formulated the rules of the deductive and inductive methods by which scientists worked, while certain ancient philosophies enriched the ways in which scientists conceived of phenomena. Pythagoreanism, for example, conveyed a vision of a harmonious geometric universe that helped form the mind of Copernicus. New concepts of natural phenomena

In mathematics the Renaissance made its greatest contribution to the rise of modern science. Humanists included arithmetic and geometry in the liberal arts curriculum; artists furthered the geometrization of space in their work on perspective; Leonardo da Vinci perceived, however

faintly, that the world was ruled by "number." The interest in algebra in the Renaissance universities, according to the 20th-century historian of science George Sarton, "was creating a kind of fever." It produced some mathematical theorists of the first rank, including Niccolò Tartaglia and Girolamo Cardano. If they had done nothing else, Renaissance scholars would have made a great contribution to mathematics by translating and publishing, in 1544, some previously unknown works of Archimedes, perhaps the most important of the ancients in this field.

If the Renaissance role in the rise of modern science was more of midwife than of parent, in the more practical realm of technology the proper image is the Renaissance magus, manipulator of the hidden forces of nature. Working within the confines of the medieval world view, engineers and technicans of the 15th and 16th centuries achieved remarkable results, which in some ways had more to do with changing the social environment than the theories of pure science. The most important technological advance of all, because it underlay progress in so many other fields, strictly speaking had little to do with nature at all. This was the development of printing, with movable metal type, around the mid-15th century in Germany. Johann Gutenberg is usually called its inventor, but in fact many people and many steps were involved. Block printing on wood came to the West from China between 1250 and 1350, papermaking also came from China by way of the Arabs in 12th-century Spain, whereas the Flemish technique of oil painting was the origin of the new printers' ink. Three men of Mainz—Gutenberg and his contemporaries Johann Fust and Peter Schöffer—seen to have taken the final steps, casting metal type and locking it into a wooden press. The invention spread like the wind, reaching Italy by 1467, Hungary and Poland in the 1470s, and Scandinavia by 1483. By 1500 the presses of Europe had produced some 6,000,000 books. Without the printing press it is impossible to conceive that the Reformation would have ever been more than a monkish quarrel or that the Scientific Revolution, which was a cooperative effort of an international community, would have occurred at all. In short, the development of printing amounted to a communications revolution of the order of the invention of writing; and like that prehistoric discovery it transformed the conditions of life. (D.We./Ed.)

Development of printing

The great age of monarchy, 1648–1789

CHRONOLOGY OF THE AGE OF MONARCHY

By the Peace of Westphalia in 1648 Calvinism had been accepted under the terms by which Lutheranism had earlier obtained recognition in the Peace of Augsburg (1555). Germany was to remain a mosaic of petty states, each with its established religion in conformity with that of its ruler. The independence of the Swiss cantons and the United Provinces had been confirmed.

It is symbolic of a new period in European history that no ruler, whether Catholic or Protestant, paid heed to the Pope's denunciation of the Peace of Westphalia. Though political upheavals continued after 1648, the modern sovereign state had emerged as the continent's most powerful institution, with little room for rivals or rebels.

Control of the state. The growth of the state had been inextricably bound up with the problem of who controlled it. Deeply rooted in European history was the principle of the contractual nature of the royal power. The need to secure consent to important actions and, concurrently, the need to get help with administrative work resulted in a blanketing of Europe with political assemblies at all levels. But the challenge to royal ambitions represented by these assemblies had caused princes to set out to reduce or destroy them, and by the mid-17th century only England and the Netherlands showed effective resistance to the symptoms of decay.

Dutch burghers and English gentlemen could not set the pattern for the 17th century. One of the first victims of the Thirty Years' War had been the estates assembly of Bohemia, which had supposed that it had the power to depose and elect kings without consulting the Habsburgs. Soon afterward in Denmark-Norway the king brushed aside his quarrelling estates and set up rigorous absolutism. At the same time the princes of Germany took up the offensive, and by the end of the century only five had failed to shake their assemblies.

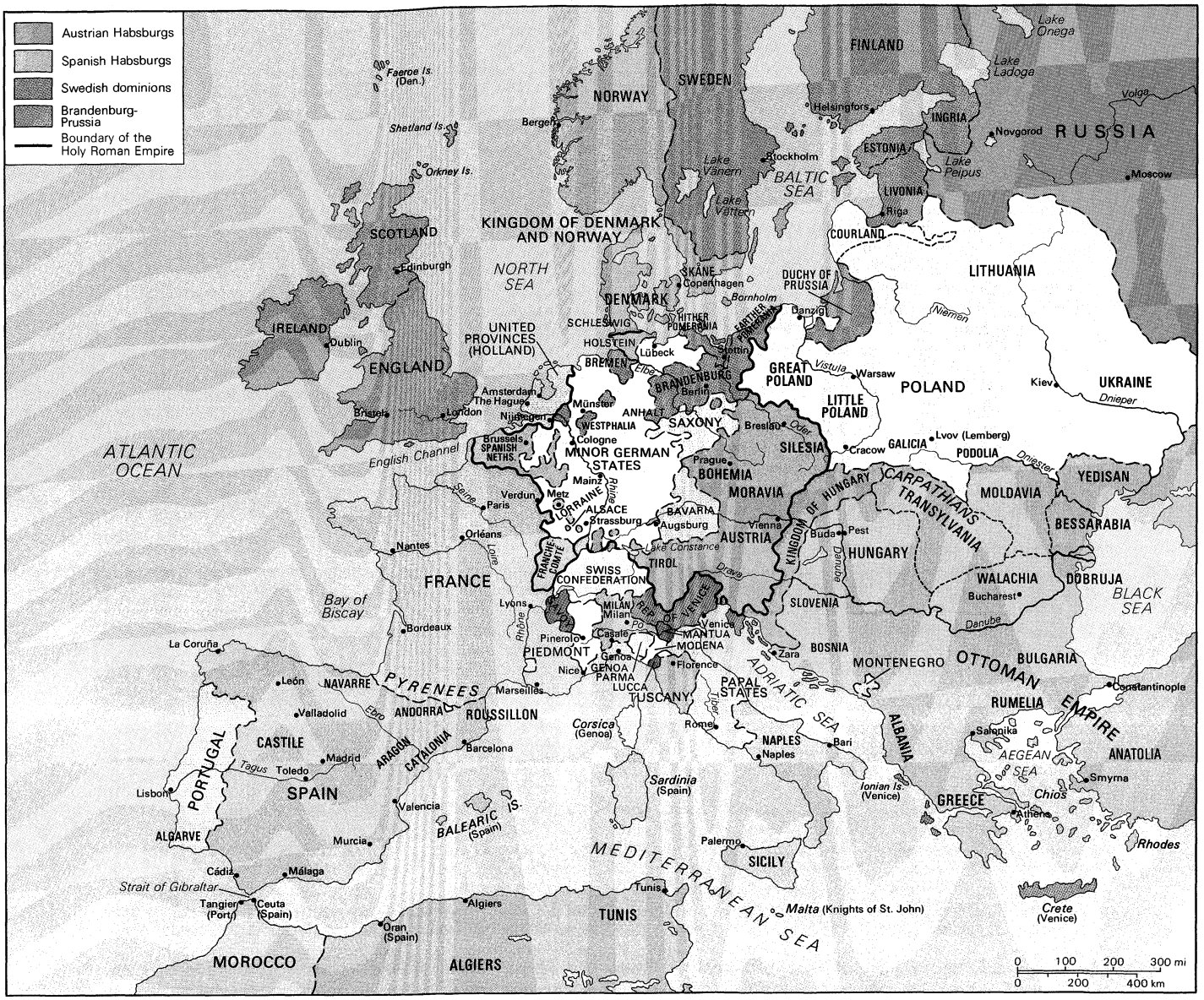
In the 18th century the estate assemblies continued their downward course. In the three continental lands where they maintained their power, they proved to be disastrously incompetent. In the Netherlands, declining economically, the parochial-minded assemblies fought the leadership of the House of Orange, and the country was headed for a revolution of its own when it was overwhelmed by that of France. Sweden, in turn, had been undoubtedly the most advanced state of Europe, but the death of Gustav II Adolf (1632) opened the door to domestic as well as international troubles. Social quarrels led in 1680 to a virtual absolutism, but the events of Charles XII's reign returned dominant power in 1719 to the Riksdag (parliament). Continuing social tensions, however, were reflected in quarrels within parliament which opened the door for French and Russian competitive corruption, and the country was saved from disintegration only by a royal coup d'état in 1772. The estates of Poland, meanwhile, had brought themselves to the point where a last-minute rescue was impossible. By the 18th century the nobles' zeal for golden liberty had reached the absurdity of the liberum veto; one dissident voice could not only block a legislative act but also "explode" the assembly. Surrounded by three hungry and militarily strong neighbours, the Polish nobles by their corruption and their divisions invited the disaster of the partitions from 1772 to 1795.

The decline of assemblies representative of estates was counter-balanced by the rise of absolute monarchy. This shift had not come about simply by default of the estates; a vital part of the story was a millennium of dynamic energy that Europe's crowned heads had thrown into the struggle for power, of which Louis XIV was the brilliant embodiment. Not very intelligent, he had, nonetheless, the talent to make his people confuse God's rule and that of the king of France. In his thirst for glory, he made his whole life a ritual of Byzantine majesty. He was a master builder whose works reflected his grandeur; he patronized the arts; he tamed disorders in his own land; and he sent his armies abroad. The lesser sovereigns found in him their model. They learned French; they copied, as far as their incomes allowed, the buildings and the manners of Versailles; and, above all, they laboured to imitate Louis's climactic achievement, his despotism.

Louis XIV

Machinery of the state. Being an absolute monarch required more than postures and parades. Success largely depended on the nature and the quality of governmental machinery. The tendency of 17th-century Europe was to assign the major tasks of government to councils or commissions. In Sweden a reorganization in 1632 had set up five ministerial colleges that were so effective that they became the model for eastern Europe and notably for Peter the Great in his efforts to reform Russia. In Spain a multiplicity of councils had stemmed from the formative days of the monarchy, and in France Richelieu adopted the Spanish style as a means of allowing his subordinates to restrain each other. Louis XIV, heir to this calculated confusion, did much of his royal business in five separate bodies. Likewise in England, both before and after the Stuart Restoration of 1660, boards and commissions were the favoured agencies. Such machinery, evolving in response to the growth of the state's business, was cumbersome, often overlapping, and prone to paralysis. As a means for coping with the resultant difficulties, the late 17th century and the 18th turned increasingly to a single minister or secretary of state. Reform not infrequently turned out to be merely an exchange of one confusion for another.

The state's supreme domestic problem was to provide a coherent direction for this inefficient machinery. The English had found an effective answer: policy formulation came from a group, or cabinet, composed of men who individually managed the great departments. The rising importance of political parties tended to bring like-minded men to office, and when George I (1714–27) absented himself from Cabinet meetings active leadership fell to a prime minister.



Europe in 1648.

In France Richelieu and Mazarin had served as over-all directing ministers, but Louis XIV found such an office incompatible with his majesty and after the death of Mazarin no one took his place. This jealousy of powerful ministers meant that the king alone not only decided policy but also carried the burdens of a chief coordinating administrator. In this respect also, Louis XIV set the precedent for the 18th century. Occasionally there was a chief minister in some capital, but the despotic temper of royalty disliked the expedient. For the same reason, royal councils as well as guiding ministers went into decline. Various attempts continued to be made in the 18th century to coordinate central functions, but royal impatience blocked the way to success. Accordingly, as continental governments faced ever-mounting obligations, the sovereign had no escape from the role of first servant of the state. In England it came to be of relatively little importance who wore the crown, but on the continent the prosperity of a country depended almost entirely on the personal qualities of the monarch.

Provincial government. Out of special circumstances federalism became established in the two republics that gained their formal independence in 1648, Switzerland and the Netherlands. Elsewhere in Europe the drive had been toward the unitary state, with immediate control from the centre. The difficulties had been enormous, because the great states of Europe were synthetic creations pulled together out of diverse territories whose institutions the king was generally bound to respect.

The English arrived least painfully at a solution. The crucial achievement of medieval English kings had been the supremacy of the law. This enabled the central government to leave local administration to the justices of the peace and the town councils. On the continent local legal variations had been too strongly embedded, and some other means was necessary for achieving unity. That means was the dispatch from the capital of a corps of loyal agents. Sweden could do this as early as the 16th century because of the small influence of feudalism in that country. Ferdinand and Isabella began in Spain by establishing their own corregidores in the municipalities, but traditional divisions of the joint monarchy made for slow progress. In France the procedure developed over centuries. The first intendants to go out from Paris were merely reporters; Henry IV and Richelieu used them as temporary agents for dealing with emergencies, and it was not until Louis XIV's reign that they became institutionalized as permanent officials superimposed upon, and therefore handicapped by, the old provinicial governments.

While Louis XIV was installing his intendants, the Great Elector of Brandenburg-Prussia began to send out war commissioners, at first to receive tax payments but soon to settle down as local officials. A second bureaucratic machine had been set up to administer the elector's patri-

monial lands, and in 1723 both were combined under one central Directorium. The directory itself was cumbersome and inefficient, but its civil service, the best in Europe, took the royal will throughout the kingdom.

In the Habsburg dominions peculiarly formidable barriers stood in the way of a centrally controlled provincial administration because of the unassimilated diversity of titles and crowns. When Maria Theresa (1740–80) undertook the creation of a comprehensive central government, she also began to move against the confusion of provincial immunities. Fairly successfully she introduced district captains as immediate agents of Vienna into the western part of her inheritance, but she hesitated to send them into Hungary, where high sheriffs drawn from the local nobility administered the counties. Her son Joseph II had no patience with such softness and sent in his German-speaking bureaucrats—to the lasting misfortune of central Europe. In these Habsburg lands above all others, some accommodation of local differences was essential, but the spirit of 18th-century despotism demanded a rational uniformity.

Maria Theresa

Government officials. In the early labours of statemaking and administration, the medieval kings had had to depend for aid on the nobility and the clergy, neither of which was single-mindedly devoted to the royal interests. With the development of urban life there had appeared as competitors of the higher orders for governmental posts members of the bourgeoisie who could offer unqualified loyalty and technical skills in law and finance. Louis XIV, who never forgot the Fronde, the nobles' revolt of his childhood, and who was jealous of any light that was not a reflection of his own, went to the extreme of closing the doors of his councils to the French grandees and transacted the state's business with "vile bourgeois." In turn the intendants carried to their offices in the provinces the same nonnoble stamp. But even Louis could not deny the spirit of the times altogether; and, following precedent, he sold to his bourgeois officials their positions and titles of nobility. This created a second aristocracy, known as the nobility of the robe because of the gown of office. The old nobility of the sword greeted the robe with a disdain that was fully reciprocated. Before the middle of the 18th century the two nobilities were uniting to oppose the remnants of Louis XIV's bourgeois policy. Increasingly arrogant about the importance of ancestors, they harried the king into giving them a virtual monopoly of the important state posts.

Louis XIV's antinoble prejudice had run aground on the stronger prejudice that nobles were the only fit companions of a king. The simplest means of meeting it was to elevate commoners as they rose in the hierarchy, and this practice became universal. In a number of countries the scale of title-making was such as to constitute, as in France, a new aristocracy. Ivan IV the Terrible had raised up a second Russian nobility committed to state obligations, but the two groups amalgamated; and Peter I the Great shortly after 1700 had the task to do over again. He attempted it in substantially the same spirit; the public servant received a status in the hierarchy of nobility in accord with his grade in the state service.

The role of the nobility

High state office, therefore, was an almost exclusive prerogative of the nobles old and new, except in England. The nobles' attitude toward their right varied from country to country. For the English gentleman public activity was a matter of personal sense of duty. The Prussian nobles, happy in their new alliance, served their king with military devotion. The Russians, resentful of service, wheedled the foolish Peter III into releasing them. The Poles constituted a special anarchistic group, and the Hungarians were potential rebels against Habsburg encroachments. The French nobility was moving into a dissident role before the death of Louis XV (1774). Through the Parlement of Paris, the chief law court, it was beginning to challenge the king's despotism.

Internal functions of the state. The growth of governmental machinery reflected the ever-broadening functions of the state. The intensified search for money led to the invasion of provincial and local government and to the endless work of tax-collecting. The maintenance of a properly majestic establishment for royalty was a state duty; though the scale varied, the style of life led to mounting extravagance in almost every land. Yet such expenditures were not to be compared to the outlays for military purposes. One of the most portentous innovations was the nationalization of the European army. By the end of the Thiry Years' War the change from older forms of organization was well advanced and the state was becoming the great consumer of the goods as well as the services and the blood of its subjects.

The oldest state functions had stemmed from the concept of the king as the fountain of justice. Concurrently with the building of a royal court system, the bitter issue of the competence of the king's courts as against traditional noble and church jurisdictions had been fought. By 1648 the major contest had been settled, although important fragments of the old order remained. Over much of Europe the nobles maintained manorial courts in which they administered justice to their peasants and serfs. The church had lost its vast legal competence but continued to bring into its own tribunals religious cases, defined more broadly than a later time would allow.

The tide continued to favour the state in its contest with the church. In Lutheran countries the state churches of the Reformation lived on without political crises. By an implied contract the princes looked after the administration, and the ministers preached submission to God and to his divinely ordained regents on earth. In England the established church was yet more clearly the creature of the state in that even its creed was a product of parliamentary legislation. Temporarily broken during the Commonwealth, restored under Charles II, and then seriously menaced by the Roman Catholic James II, Anglicanism was saved by the Glorious Revolution of 1688, and it gratefully put itself at the service of the new order. During their ascendancy in the 18th century the Whigs took over the church as well as the state, and the bishops in the House of Lords dutifully supported the government.

In Russia the outcome was even more devastating for the authority and the moral position of the church. With a patriarchate of its own after 1589, Russian Christianity had tried, in the face of state encroachments, to argue for dual, independent spiritual and secular realms. An old tension came to the breaking point when the patriarch Nikon, elevated in 1652, introduced modifications in ritual for political ends. The government's quarrels were not with the reforms but with Nikon's arrogance, and the Tsar dismissed him. The church was split, but neither branch was prepared to accept the Westernizing infamies of Peter I the Great. He responded to opposition by abolishing the patriarchate in 1721 and by putting the church under a Holy Synod. Soon a lay procurator became the president of the synod and the controversy was resolved for the life of the tsarist regime: the Russian church was little more than the spiritual arm of the secular order.

The Russian patriarchate

In Roman Catholic territories the state, of course, had to reckon with a powerful international organization. The papacy, shaken by the Reformation to its foundations, had been able to restore and even strengthen its position. A proud pope and a Louis XIV believing in divine right were bound to collide. Louis's extension to the whole kingdom of his right to control the affairs of vacant bishoprics precipitated a new chapter of the conflict. In 1682 an assembly of the French clergy, prodded by Louis, stated in four points the so-called ancient liberties of the Gallican Church and reasserted the principle of the subordination of the papacy to a general council. The Pope again protested vigorously, and Louis threatened to set up an autonomous national church. For 10 years relations between Rome and Versailles were suspended. At last, however, Louis's embroilment in war led him to the motions of capitulating to a new pope. The controversy involved no suspicion of heresy; at its height Louis repealed Henry IV's edict of toleration of Protestants.

Louis's zealous orthodoxy confused and then transformed the whole church-state issue in France. He was so hostile to the doctrines of Jansenism that he appealed to the Pope for support against a vigorous segment of his own subjects. Papal bulls of condemnation inspired an alliance of

Jansenists and nationalist Gallicans, and their resistance to combined royal and papal pressures stirred a discord that lasted until the eve of the French Revolution and that did great damage to the spiritual force of official Catholicism. Anticlericalism and Deism were the end products.

While France's religious problem was taking this special turn, in the other Catholic countries the debate remained centred on the old issue of the papacy and its powers. A stream of books attacked the allegedly swollen pretensions of Rome. The most noted of them came in 1763 from a Catholic bishop, J.N. von Hontheim, writing under the name of Justinus Febronius. Febronius would allow the pope only the first honour among equals; the church's council stood above him, and the time had come to restore the bishops to their usurped position. No less vigorously did he champion the rights of secular princes in the external affairs of the church, and he invited them to suspend obedience to Rome if the pope proved adamant. The clerical opposition within the church and a whole phalanx of Catholic rulers found in Febronius practically every known justification for their ambitions and their encroachments. In 18th-century Naples and Sardinia, in Portugal and Spain, the kings extended their powers of church nominations and other functions. During the peace negotiations in 1713 at Utrecht, the state delegations excluded the papal nuncio altogether and set a precedent for the rest of the century by ignoring Rome's claims to overlordship over various Italian territories.

Nowhere was the papacy more disappointed than in the Habsburg dominions, where, since the Reformation, the ruling family had vigorously supported Catholicism. Maria Theresa was personally dedicated to the church, but as sovereign she was sensitive to the slightest hint of papal intervention. Joseph II elevated Febronianism into Josephism. He closed monasteries, sent theological candidates to Rationalist state seminaries, and even decided matters of church ritual. Joseph II, like Louis XIV, lost the support of his people, and his brother Leopold II, succeeding him in 1790, had to salvage what he could by abandoning Joseph's excesses.

The Jesuit controversy

Meanwhile the Society of Jesus had come under disastrous fire. Some of the Jesuit doctrines and tactics had inspired criticisms within the church and even rebukes from the papacy. Certain Jesuit political teachings and activities had created ill will among secular authorities. Catastrophe began in Portugal, a country the Jesuits had dominated for a century, and it came at the hands of one of their students, the reforming Marquês de Pombal. Making use of difficulties in Paraguay and alleged complicity in an attack on the king, Pombal expelled the Jesuits from Portugal and its colonies in 1759–61 and for a decade suspended all ties with Rome. In France, Jansenism, Gallicanism, and anticlerical Rationalism united in a long-sustained attack brought to its climax by the bankruptcy of a Jesuit commercial firm. The Paris law court found the society subversive and ordered its expropriation and expulsion (1764). Three years later the Bourbon king of Spain drove out nearly 6,000 Jesuits, and shortly Naples and Parma followed suit. The Bourbon sovereigns then jointly demanded the suppression of the society, and in 1773 Pope Clement XIV, a Franciscan little disposed toward the Jesuits, acquiesced.

In addition to these concerns with the administration of the church, the state also had the traditional function of defense of correct Christian belief. The Reformation had given the rulers of Europe an opportunity to decide, each for himself, what was correct, and the starting assumption was that subject would bow to the decision of sovereign. It had been increasingly difficult, however, to impose a single form of faith. In the Low Countries a formal Edict of Toleration in 1614 arose out of very practical economic and political considerations. In England shifts of official religion from Henry VIII's break with Rome to the Glorious Revolution had strewn the land with a diversity of churches; and the Act of Toleration in 1689 simply regularized an inescapable reality. In Germany the Peace of Westphalia had allowed the princes of the Holy Roman Empire to impose on their subjects their own choice among three religions. In the regions devastated by the Thirty Years' War there was acute need for population increase, and the Great Elector of Brandenburg-Prussia was more interested in an immigrant's ability to work and pay taxes than in the minutiae of his Christian convictions. In the next century the personal indifference of Frederick II the Great gave new force to old policy; as long as the churches supported the monarchy, he was content to allow freedom of religious expression.

The Roman Catholic states lagged behind this Protestant march toward toleration. Louis XIV's economically injurious resolution in 1685 to destroy Protestantism was unrevoked for 102 years, until weariness with religious quarrels, the undermining thrusts of Rationalism, and the commanding power of secularization forced Louis XVI to restore a limited tolerance. In Austria, Maria Theresa saw in tolerance only sinful indifference; but Joseph II, who had no patience with his mother's anxieties, issued the Patent of Tolerance in 1781, which gave to non-Catholics a measure of religious freedom and an improved civil status. This caprice of his religious policy brought the Pope himself on a protest mission to Vienna, but to no avail.

The state in its external relations. The Peace of 1648 revealed in unclouded sharpness the sovereign independence of the European state. The old claimants to universal authority, the empire and the papacy, had withered into impotence, and by recognized right the state defined its own interests and decided how best to protect them.

For the conduct of peaceful international relations, mechanisms and procedures had been developing since Renaissance days in Italy. The basic institution was the foreign ministry, which emerged out of the administrative uncertainties of the major continental capitals around 1700. England lagged behind; two secretaries of state divided external affairs until 1782. But the evolution of a central agency was no guarantee of system and order. Foreign affairs always had been a jealously guarded prerogative of the crown, and policy decisions normally remained subject to those calculations and accidents that moved a king. The second instrument of diplomacy was the ambassador, by custom a man of high birth. Long experience had defined the ambassador's duties and conferred on him privileges and immunities. The charge to report significant developments gave him the aura—and the functions—of a spy, and the obligation to protect his sovereign's honour made him truculent about the honours due to himself. About 1500 the papacy had devised a strict order of precedence for all emissaries in Rome, but toward 1648 this scheme of relative merit was no longer acceptable to Europe's monarchs, filled as each was with a sense of personal majesty. Fierceness about questions of rank made it necessary to hold two peace conferences in Westphalia, and at each the delegates bickered over places in ceremonial processions and even over how far a host should accompany a retiring guest toward his carriage. Forethought and practical expedients eased some of the tensions in the 18th century, but only later times fully solved the problems in terms of the equality that proud kings had so long demanded.

Military establishment. The organization of a satisfactory military establishment proved a difficult task. The feudal system still existed, although largely on paper, in several parts of Europe until about 1700, but for centuries it had been of little value. The state groped toward a better solution when it began to create an army of its own. There were a few standing companies in France by 1500 and more in 16th-century Spain, but the senior regiments of Prussia and Austria dated from the Thirty Years' War. A fully state-owned and operated army came only slowly in the 18th century.

For the officers' corps tradition designated nobles, but sons of the bourgeoisie had moved in when the aristocrats showed incompetence in the skills of artillery and engineering and when, in some armies, commissions were sold as private property. In England this intrusion was merely disliked, but on the continent resentments grew as the lines of social exclusiveness hardened. In France the nobles in 1781 at last wheedled out of Louis XVI an edict limiting commissions to men able to prove 16 quarterings of nobility. Meanwhile they had gained another victory:

the multiplication of posts until the corps became a vast program of relief, and the higher commands had to be rotated frequently to give a semblance of duty. In Prussia the nobles also enjoyed a strong preference, but the results differed markedly from those in France. Instead of granting demoralizing sinecures, the Prussian kings demanded a total dedication to duty and paid for it by recognizing in their brothers-in-arms the foremost order of the realm. In the other European armies the prevailing aristocratic stamp suffered little damage.

The nationalized army

The common soldier came from, and lived in, another world. The nationalized army began under the theory of volunteer long-term enlistments, but very early free decision did not produce enough soldiers. Public policy disapproved of taking useful workers, and other sources had to be tapped. Enlisting foreigners proved an effective resort; their proportion in the continental armies ran at times from one-quarter to two-thirds. Within the country the preferred solution was a roundup of the economically nonproductive. Local authorities collected jailbirds and paupers and misfits, while recruiting sergeants took in by tricks, and if necessary by violence, the innocent and the unwary. The inadequacy of such devices gradually forced the state into instituting compulsory service. Sweden was an innovator in a systematically operated draft, but the more momentous step occurred in Prussia around 1730: a regularized conscription based on the principle of universal liability. A generation later Joseph II took the same course. The inclusion of peasants and artisans, however, remained a disagreeable necessity; until the French Revolution Europe preferred for its armies the local tatterdemalions and the foreign adventurers. No one had any illusions about the quality of this raw material, but ways had been devised to cope with it. The first was discipline so rigid and brutal that the soldier had to fear his officer more than the enemy. The second was drill and yet more drill until the men became cogs in a precision machine. Even so, in tactical operations little was possible beyond marching onto a field and firing volleys; any break in the ranks usually resulted in wholesale desertions.

Strategy was limited. An army was far too expensive to be thrown recklessly into battle or taken far from its base of supplies, since foraging meant decamping. Supplies, in turn, meant transport, and therefore a campaign had to close at the onset of autumn. This sense of limited objectives both in battles and campaigns meant that in the 18th century no one thought of a total defeat or of exacting severe penalties from a fallen enemy. War, whatever its dimensions, nonetheless continued after 1648 to be a ready resort of state policy. It had its critics as well as its glorifiers. In the 17th century Henry IV's great minister Sully published a project attributed to the king of a "great design" for the reconstruction of Europe and the creation of an international council to restrict warfare to Crusades against the Turks. In the generations between Sully and Immanuel Kant, men repeatedly issued their own impassioned versions of the great design for an organized peace; the dream would not die, but the reality of those generations paid it no attention.

The dream of an organized peace

Balance of power. The best device for peace and protection seemed to be the balance of power. The principle, as old as Western culture, had reappeared as conscious policy in Renaissance Italy and had moved to transalpine Europe about 1500. Its need had become apparent when the emperor Charles V inherited so many dominions on all sides of France. War resulted, and the conflict went on for so long that Franco-Habsburg hostility seemed unending and inevitable. Finally the Peace of Westphalia ended any conceivable menace to France from the German Habsburgs; and the Peace of the Pyrenees in 1659 won France some territorial advantages and a royal marriage from the Spanish. These settlements confirmed a shift in the directon of French external ambitions from Italy to the northeastern frontier.

Louis XIV began his personal rule in 1661. The threat to the balance that then appeared was of Louis's own making, and one coaliton after another rose up to resist his aggressions in the War of Devolution (1667–68), the Dutch War (1672–78), and the War of the Grand Alliance (1689–97). By the time of the Peace of Rijswijk in 1697, Louis's drive had been fairly effectively checked, but the death of Charles II in Spain in 1700 reopened the struggle. Obliged to designate an heir to his many crowns, the childless king Charles in his last days chose a grandson of Louis, his brother-in-law Philippe, duc d'Anjou. Louis's greedy response indicated the greatest upset to the balance of power since Charles V, and the widespread War of the Spanish Succession ensued.

This multiplication of battlefields was the work of England. In the 16th century imperial friction and religious animus had made Spain the great enemy; in the 17th century commercial competition had made the Dutch the prime enemy of both Oliver Cromwell and the Restoration monarchy. As Louis XIV began his depredations, he bought the support of his cousin, the English king Charles II, but on the question of France, as on so many others, Charles and his subjects drifted apart. Religious feeling, economic and colonial friction, and the French threat to the Low Countries all indicated to English opinion the emergence of a new enemy. In 1688 the Glorious Revolution placed on the throne Louis's most implacable enemy, William III of Orange, and before a year was out the new king took England into his old struggle. Thus began that Anglo-French contest that went on until Waterloo in 1815.

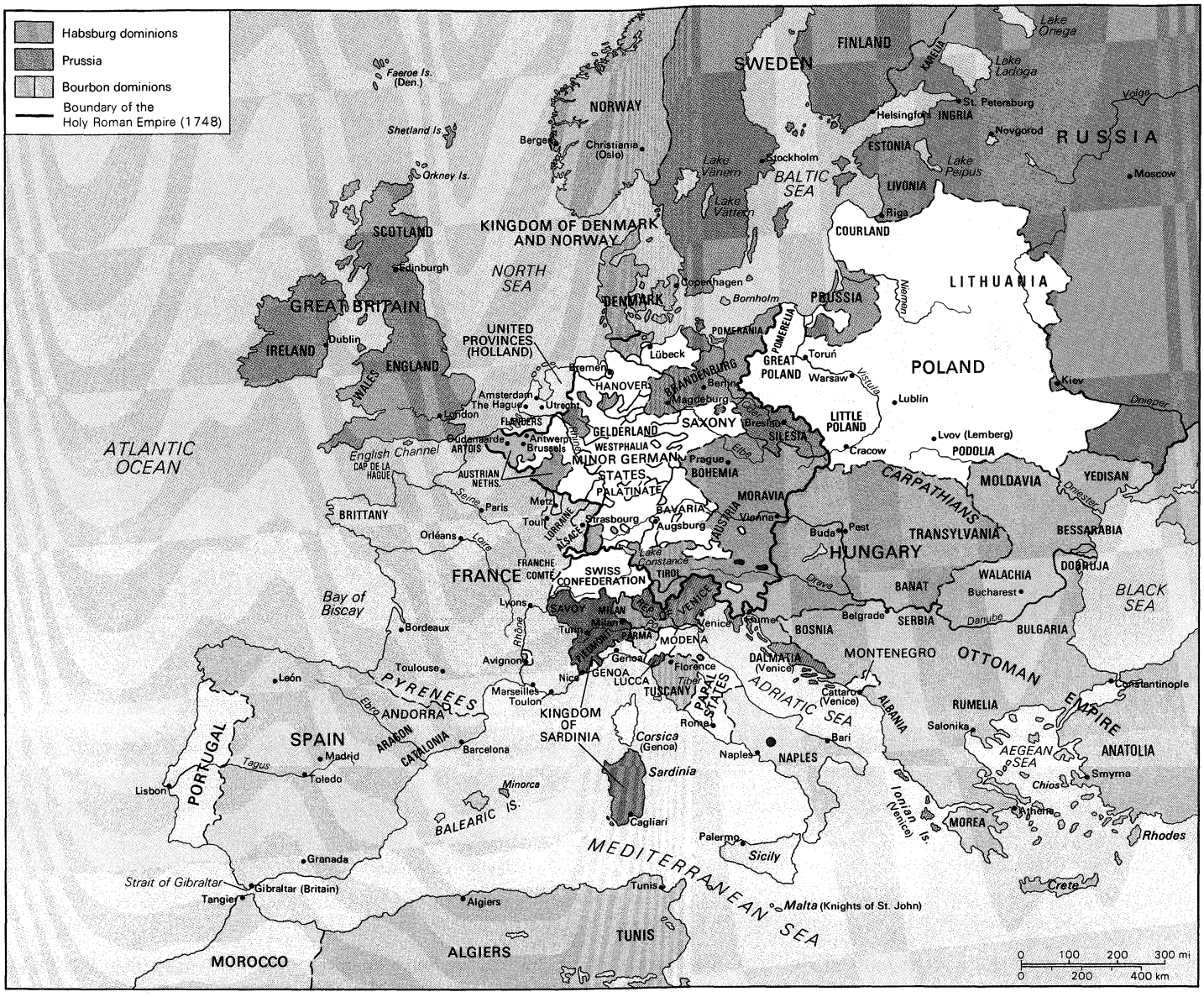
The Glorious Revolution

When exhaustion and allied cross-purposes brought the War of the Spanish Succession to an end at Utrecht in 1713, the treaties proclaimed a "just equilibrium of power." Historically conceived, as a deterrent to war, the principle of the balance as invoked at Utrecht determined the distribution of territorial compensations at the end of a war. Repeatedly thereafter until the final partition of Poland (1795) numerous crises arose from demands to restore the balance in this manner.

The Treaties of Utrecht introduced a period of cooperation between Britain and France, but unresolved imperial rivalries allowed it only a passing term. Deteriorating relations led first to the War of Jenkins' Ear between Britain and Spain in the West Indies in 1739 and shortly to world-wide conflict between Britain and France. Meanwhile Frederick II the Great of Prussia in 1740 had precipitated a second European war, the War of the Austrian Succession, by overrunning Silesia at the expense of the empress Maria Theresa. For hotheads in Paris this was an opportunity for a new attack on the Habsburgs, and soon French military successes in Germany warned of a dangerous upset to the balance of power. British interest required assistance to the empress, and the two wars intermingled. In 1748 general weariness forced the Treaty of Aix-la-Chapelle. The British had tired of the Austrian cause and Frederick kept Silesia, but otherwise the peace terms required mutual restoration of conquests. (Da.H./Ed.)

The "Diplomatic Revolution." In the years immediately following Aix-la-Chapelle, the only force that seriously threatened the peace was the studied resolution of the Austrian chancellor Wenzel Anton, Graf von Kaunitz, to win back Silesia. To be sure, an undeclared war between England and France was being fought in America (the so-called French and Indian War), but, far from desiring to have it spread to Europe, England shrank from that prospect. The English were fearful that in the event of war on the Continent either France or France's ally Prussia might overrun the Duchy of Hanover to which the English king held title and, also, perhaps occupy the Netherlands. As England's own ally, Austria, could not give adequate assurances of military aid, England turned to Russia. By the Subsidy Treaty of September 1755 England purchased the services of 50,000 Russian soldiers, who aligned themselves on the border of East Prussia, poised to advance into Prussian territory. The English did not regard these arrangements with Russia as incompatible with friendly relations with the Habsburgs. The arrangement was designed to give England double assurance.

Developments belied intentions. Prussia saw in the arrangements not a defensive move but a threat. Frederick already knew of a secret Russian–Austrian understanding against Prussia. His sense of security gave way to a deep fear of an unwanted war against Russia. The way out, he reacted in his fear, was to undercut Russia



Europe in 1748.

by offering England the assurances concerning Hanover. Since Prussian troops would be appreciably closer to Hanover than the distant Russians, England readily accepted Frederick's proffered assistance. The two states accordingly signed the Convention of Westminster (January 1756), in which Prussia agreed to deny entry into and passage through Germany of any foreign troops (*i.e.,* Russians). England obtained the desired guarantees, and its relief was great. Frederick's was equally great. He saw in the Convention an instrument to keep the peace.

The Convention of Westminster

Westminster, however, did not produce the desired effect. It was a bombshell. Vienna was outraged, or professed to be, by the "treachery" of its English ally. Louis XV and Versailles bristled over the seeming defection of Prussia, and, fearing an English–Prussian combination, France turned toward Austria as a shield.

France contracted a defensive alliance with Austria, the First Treaty of Versailles (May 1756), by which France agreed to come to the assistance of Austria if the Habsburgs were attacked. To defend Austria, so French calculations ran, was also to protect themselves. Thus, out of this reflex of fear, France took the first long step toward a reversal of alliances and the abandonment of its centuries-old anti-Habsburg policy.

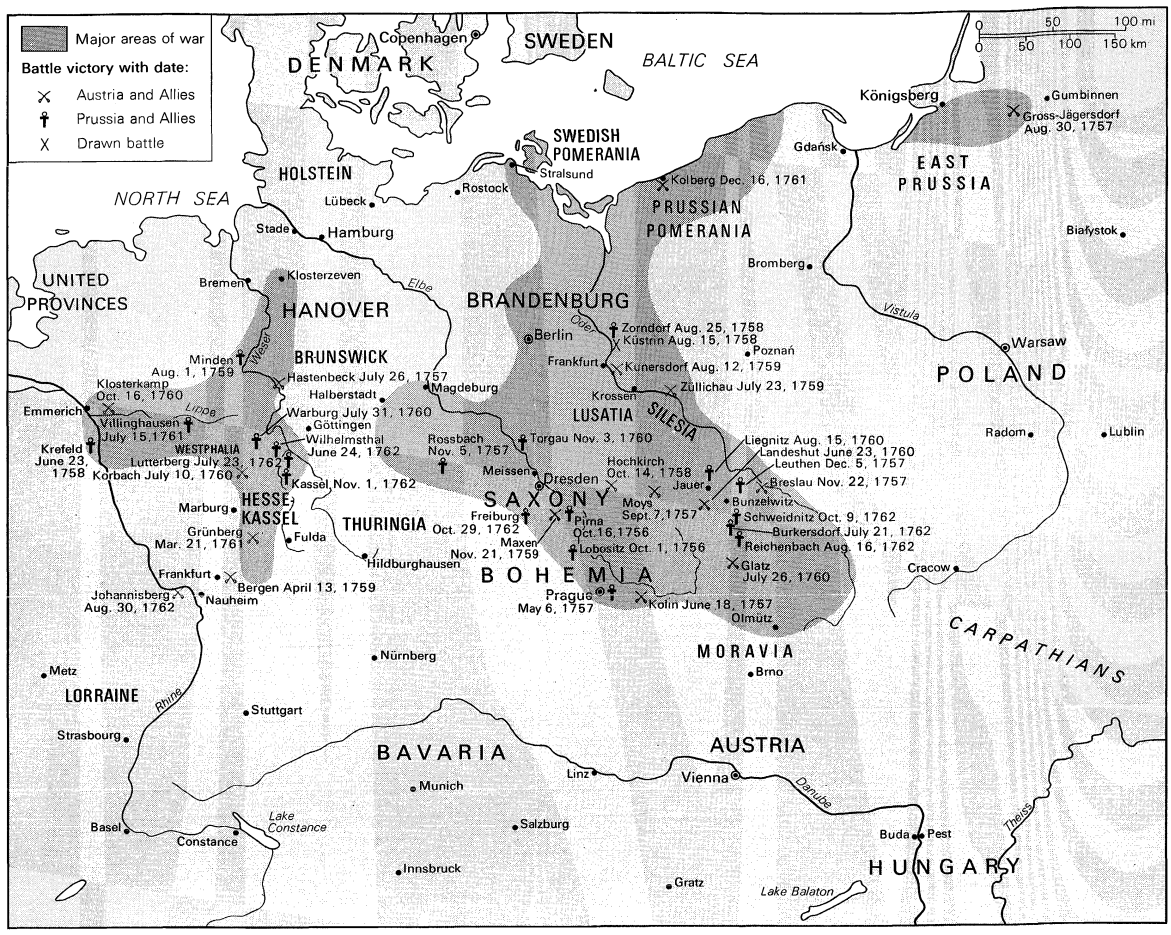
Reversal of Bourbon–Habsburg enmity

St. Petersburg's reaction was, if anything, more violent still. To Russian eyes, Westminster was clearly an unscrupulous repudiation of earlier English policy. Thus, a third set of fears obtruded on the diplomatic scene—Russia's fears that it might have to fight alone against Prussia covered by England. Russia thus turned to Austria for self-protection; in May 1756 Russia agreed to give Austria limited military aid in the event of a Prussian attack.

Frederick did not wait to be attacked first. In August 1756 he invaded Saxony, which he considered the centre of plotting against him. It was now too late for diplomacy: the generals had taken over. France contracted a full military alliance with Russia in the fall of that year. In the following spring, before the campaign began, France signed the Second Treaty of Versailles with Austria, a full offensive–defensive military accord. By this agreement, the last of the three bilateral pacts, the reversal of alliances had become a *fait accompli.*

This diplomatic revolution was a lesson in futility, revealing the fatal inadequacy of the balance of power. Seeking peace, the great states were led to create a situation beyond their control. Obsessed by reciprocal mistrust, they made the appeal to arms inescapable.

The Seven Years' War and the Peace of Paris. Frederick's invasion of Saxony opened the war on the Continent, which merged with the commercial–imperial conflict into the Seven Years' War (1756–63). In the European theatre, the Austrians at once invoked the provisions of the mil-



Main areas and sites in central Europe associated with the Seven Years' War.

Adapted from H.C. Darby and Harold Fullard (eds.), *The New Cambridge Modern History*, vol. XIV, Atlas (© 1970), Cambridge University Press; maps © George Philip & Son, Ltd. 1970

itary agreements that bound them to France and Russia. Frederick, for his part, demanded and obtained English aid in the form of large subsidies and a small task force dispatched to the Continent. French tactics envisaged a rapid campaign for the conquest of Hanover. Military calculations in England were initially divided, veering after Frederick's great victories at Rossbach and Leuthen in 1757 toward a full military understanding with Prussia and the use of Hanover as the base of operations for a large-scale military diversion to contain the enemy.

Frederick's stubborn defense

Frederick's conduct of the war was heroic for its daring, flexibility, and dauntless courage against overwhelming odds. It was victorious, too, for a brief span of years. Operating on inner lines until 1759, he prevented the Austrian and Russian armies from joining forces. In 1759 the Russians inflicted a shattering defeat upon him at the Battle of Kunersdorf, and his future seemed hopeless. Yet somehow he held on, suffering defeat after defeat but refusing to sue for peace. For three years the war in Europe was a protracted and ruinous combat in which the Prussian and Austro-Russian armies confronted each other in the east, while French and Anglo-Hanoverian troops effectively neutralized each other in the west. In Prussia's exhaustion Frederick could not have resisted indefinitely, particularly since England withdrew its financial support, but on January 5, 1762, "the miracle of the House of Brandenburg," as patriotic Prussian historians call it, or "her sacred majesty, chance," as Frederick put it, saved him. Tsarina Elizabeth died and was succeeded by her deranged nephew Peter III, an unmeasured admirer of Frederick. With Peter's accession the anti-Prussian coalition fell apart. After signing immediate peace with Prussia the Tsar restored all Russian conquests and placed the Russian army at Frederick's disposal. Though Peter was assassinated within a half year, Peter's wife and successor Catherine II the Great wanted no renewal of a war in which her country would be fighting against Austria for the benefit of Prussia. Prussia's eastern front was at last secure. "Thanks to providence," Frederick jubilantly cried out, "our back is free."

The major war in the Germanies now virtually ceased, but the fighting for empire continued on the seas in America, the West Indies, Africa, and India. The British conquests outside Europe propelled William Pitt the Elder, Britain's war leader, into the pages of history as the great empire builder. On the other hand, in France the dictatorship of Étienne-François, duc de Choiseul, not less embracing and intransigent than Pitt's, failed to win security for France. With rising opposition at home, defeats on the high seas, and the failure of a direct invasion of England and Scotland, Choiseul made overtures for peace early in 1760, not without first playing the Spanish card in his negotiations. Fortunately for England, Pitt's fears of Spanish might were exaggerated. Spain suffered crushing defeats on all fronts in its brief military venture.

After these reverses nothing remained but for Choiseul to sue for immediate peace, which was signed on February 10, 1763. The end of the war between England and France did not come, however, until John Stuart, 3rd earl of Bute, the chief English negotiator, had already taken his country out of all fighting on the Continent. He broke completely with Prussia, cutting off all subsidies and leaving his erstwhile ally to its own resources. Austria, too, was then ready for negotiations, for after Catherine's withdrawal of Russian troops the Habsburgs stood alone in Europe.

The Treaty of Paris was less severe than might have been expected in view of Britain's overwhelming victories. The peace was the handiwork of Lord Bute and of still lesser men who had forced Pitt from office and were prepared, to secure their advancement, to sacrifice if necessary some of the imperial interests that Pitt had so tenaciously and energetically pursued. The guiding consideration of Bute was to turn the diplomatic negotiations to the account of his young royal master, George III. The tactics of the

imaginative Choiseul, the chief French negotiator, were to drive a wedge between the British peacemakers and exploit to the full their fear that Pitt might be returned to office and reopen the war. For Choiseul the peace was to be only a respite from actual fighting, an armistice so to speak, which France would terminate when it was again prepared to reopen the duel for overseas empire. Though he made bitter concessions to the victor, he saved France from the status of a second-rate power, which Pitt had planned. Though Britain did not obtain the Carthaginian peace that Pitt would have imposed, it had emerged as the leading commercial and colonial power in the Western World.

The emergence of Britain's colonial supremacy

In North America, Britain became master of the vast territory east of the Mississippi as well as "the frozen acres" of Canada. Patriots angrily clamoured that their country also retain the captured West Indies sugar islands; however, despite the great strategic naval importance of Guadeloupe and Martinique, the sugar bloc in Parliament sacrificed imperial interests to financial profits and voted their return to France. Moreover, France was permitted to keep the island of Gorée, commanding the slave trade of Senegal, along with a number of trading posts on the eastern coast of India.

Spain, too, fared better by the peace terms than might have been expected by virtue of its overwhelming defeat. The Spanish lost Florida to England and gave up their old claims to the Newfoundland fisheries, but they kept their South American colonies and islands in the West Indies. The victor restored Cuba and the Philippines to Spanish rule. Choiseul, in order to place a lien upon his Bourbon ally's goodwill, ceded to Spain that part of the French territory of Louisiana, including the port of New Orleans, which lay west of the Mississippi.

Prussia also ended the war a victor, for the separate Treaty of Hubertusburg (February 15, 1763) formally recognized Frederick's conquest of Silesia. In Prussia there was great devastation and suffering, but that did not last. What lasted was "the miracle of the House of Brandenburg," the survival of Prussia, its new status as a great power, and the continuing confrontation of Prussia and Austria for mastery of Germany.

Renewal of Anglo-French competition. Choiseul was the directing intelligence behind his country's foreign and colonial policy during the 1760s. He counted heavily on the Habsburg alliance of 1756 and on the network of Habsburg–Bourbon dynastic marriages to cover French security on the Continent and leave France free for action outside Europe. Working via the Pacte de Famille with Spain, Choiseul looked ahead toward a joint Bourbon thrust against England and bent every energy toward restoring the shattered might of French armed services and strengthening the bases for naval operations. This aggressive policy could not be implemented, however, and under Charles Gravier, comte de Vergennes, who took over in 1774, it was toned down. On the Continent Vergennes relied less on the Habsburg alliance (which he feared would entangle France in Austrian intrigues) and more on developing friendly relations with Prussia, with the German states along the Rhine, and, most of all, with Russia. There was appeasement in this approach, but he was prepared to pay the price of lessened French prestige and influence to obtain economic benefits. "In the present state of affairs," he declared, "commercial questions are political questions and as such are within the province of foreign affairs."

The U.S. War of Independence

Vergennes's great opportunity to strike directly against England came in 1776 with the armed resistance of the American colonies to British rule. Initially, he adopted a policy of open non-belligerency against England and more or less secret assistance for the Americans. The decisive change from secret aid to open alliance came after the U.S. Declaration of Independence in July 1776 and Gen. John Burgoyne's defeat at Saratoga in 1777. In February 1778 Louis XVI signed treaties of alliance and commerce with the Americans, and within a few months France was openly at war with England. Vergennes had not miscalculated on the Bourbon ally. In 1779 Spain contracted an open alliance, contributing with the United States 40 ships of the line that gave the associates naval superiority. In return, Spain obtained from Vergennes a reluctant agreement not to withdraw from the war until Gibraltar had been retaken from the British.

In 1780 England declared war against the United Provinces, who had built up a lucrative trade conveying supplies to the colonists and allowing American privateers to refit in Dutch ports in both Europe and the West Indies. But Vergennes appealed to the neutrals to encompass England's defeat by turning them against "the tyrant of the seas." Here his warm relations with Russia stood France in good stead, and Russia was persuaded to sign the declaration that established the League of Armed Neutrality. By the terms of the declaration, contraband was narrowly restricted to munitions and arms, and the armed neutrals regarded themselves as free to navigate along the coasts and to the very ports of all belligerents. Vergennes had gone far toward arraying the Continent against England.

The peace treaties of 1783, signed by England, France, Spain, and the Netherlands, ended the fighting. The treaties recognized the independence of the colonists. Spain retained Minorca and Florida. To France England ceded Tobago and Senegal. England retained bitter memories of a long and humiliating series of naval reverses. Though Adm. George Brydges Rodney's victories in 1782 restored English superiority, the antiwar mood of shame and humiliation was deep. The English wit Horace Walpole caught it in a striking phrase: "You must be happy now," he wrote, "not to have a son who would live to grovel in the dregs of England." England's enemies were jubilant, more jubilant than reality warranted. "Thus is the great power . . .fallen completely and forever," wrote Leopold of Tuscany, " . . . sunk to the ranks of a second class power . . . . France, delivered for all time of that formidable adversary, thereby doubles her intrinsic strength, her commerce, her authority, and her prestige . . . ."

The growth of French commerce

France did not double its "intrinsic strength," least of all in the military affairs of eastern Europe. With respect to the French commercial position between 1783 and 1789, however, Vergennes's policy did add greatly to the strength of the country. He negotiated a trade treaty with Russia that gave French merchants terms as good as those given to the English. By opening trading facilities to foreigners in its West Indies colonies (1784) and concluding a practically free-trade treaty with England in 1786, France accelerated the remarkable increase in the volume and value of its commerce. And England was far from "fallen completely and forever." By skillful diplomacy England formed a triple alliance with Prussia and the United Provinces. As an ally of the United Provinces, England was assured of navigation on the Scheldt and a point of entry for trade with the Continent. As an associate of Prussia and in possession of Hanover, England could build a trading clientele in the Germanies. A dominant member of the alliance and Russia's greatest customer, England pursued active trade relations in the North Sea and the Baltic. With Gibraltar in their hands and with the support of Portugal, the English also held Spain in check. By supporting the territorial integrity of the Porte (Ottoman government) they had access to the eastern Mediterranean, and their share in the trade with free America surpassed the total reached before 1776. On the eve of 1789 the two maritime states still confronted each other in the competition for empire. (L.Ge.)

Eastern Europe and the Eastern Question. A new period in eastern Europe was also opening toward the middle of the 17th century. Poland, once stretching over the vast plains from the Baltic to the Black Sea, was beginning to pay the price of its consitutional follies as one neighbour after another carved for itself slices of the country. In contrast Sweden entered the political scene with its drive to dominate the Baltic. High moment and disaster came in rapid succession after Charles XII came to the throne in 1697. He struck at foes from Denmark to Russia, but in his second campaign against Russia—in the Second, or Great, Northern War—he marched deeply into southern Russia to join the Cossack Ivan Mazepa and his army and suffered annihilation at Poltava in 1709. Charles was one of the few to escape, but his glory had departed. The Treaty of Nystad, made in 1721 after his death, merely

confirmed the fact that Russia had taken the eastern shore of the Baltic, where already the tsar Peter I the Great was building his new capital of St. Petersburg. Charles by invading the heart of Russia had opened the way of Russia into Europe. Russia was also in those years expanding southward. In 1677 the Russians stumbled into their first encounter with the Ottoman Turks, and this scuffle launched a new theme of European history. Peter took up the challenge and by 1697, after bitter reverses, had fought his way down to the Sea of Azov. Already a Russo-Turkish war was caught up in a broader context. Habsburg armies were also on the march against the Ottoman Empire, and the consequence was a question laden with implications: how were these enemies of Turkey going to deal with each other?

In 1683 The Turks began their second Siege of Vienna; and, when King John III Sobieski appeared to relieve the city, Ottoman power collapsed. In pursuit, the Christian forces pushed down the Danube to liberate most of Hungary. The eternal necessity of Austria to face both East and West interrupted the work—Louis XIV was making trouble—and it was not until 1697 that the emperor Leopold I was ready to strike a decisive blow. Meanwhile Peter I the Great, desirous of promoting a Grand Crusade, had sent an emissary to Vienna, and there the two enemies of Turkey signed a three-year alliance. The Austrians inflicted a crushing defeat on the Sultan at Zenta; but then, to the consternation of Peter, Leopold insisted on making peace; he had to prepare for the death of the King of Spain. In the Treaty of Carlowitz (1699) the Emperor recovered Hungary, and the Tsar had to content himself with Azov. This first essay in cooperation, therefore, turned out badly. In 1710, when the Sultan declared war on Russia, Peter proposed a new alliance; but the War of the Spanish Succession was absorbing all of Austria's energies, and he had to fight alone. Suddenly caught by the Ottomans on the Pruth River, Peter bought his escape by surrendering Azov; when, a few years later, there came an Austro-Turkish explosion the Tsar remained at home. There was again, consequently, a separate war, but this time, after Eugene of Savoy had won more brilliant victories, Austria at the Peace of Passarowitz (1718) gained possessions in Bosnia and Serbia including the citadel of Belgrade.

After Peter's death (1725) the ruling clique in Russia rekindled the anti-Turkish fires and, in preparation for fresh conflict, negotiated a defense treaty with Austria. When in 1735 the Russians were the aggressors, Austria was not bound to fight, but Vienna discovered a powerful argument for war: there would be great danger in allowing Russia to fight its way down to the Danube and there dictate a private peace; cooperation in the war would ensure cooperation in the peacemaking. Calculations and hopes alike went awry. There were bickerings between the allies over their uncoordinated efforts; and, just as the Russians had won a solid victory in Moldavia, panic seized the Austrian army. Amid great confusions an emissary signed the Treaty of Belgrade (1739), which returned to Turkey almost all of the gains of 1718. Unable to carry on alone, the Russians went home, and St. Petersburg made peace with the small comfort of recovering Azov.

For a generation the two former allies had no time for Turks, but in 1768 Catherine II's activities against the Poles set off a new Russo-Turkish conflict. Within a year Russian troops stood on the Danube and, in gravest alarm, the Austrians made an alliance with the Sultan, pledging themselves to rescue Poland and to defend the territories of Turkey. Instead of fighting, however, Joseph II and Maria Theresa turned around and joined Russia and Prussia in the first partition of Poland (1772). Catherine was then free to settle alone with Turkey, and in the Treaty of Küçük Kaynarca (1774) she won spectacular territorial and commercial gains and, most ominous of all, secured a vaguely defined right to make representations on behalf of the Greek Orthodox Christians of Turkey.

Joseph II had walked out on his ally but justified himself on the ground that he had moderated Catherine's demands. It seemed to the emperor better to work with the empress than against her, but over several years of negotiations they could not agree on objectives. When Catherine in 1787 provoked a new war, Austrian forces, nonetheless, also took to the field. Neither army fought well, however, and suddenly the new king of Prussia, Frederick William II, struck at them by allying with Turkey. In the midst of these and other misfortunes the dispirited Joseph II died and his brother Leopold II, abandoning the Russian accomplice, made peace at Sistova in 1791 on the basis of the *status quo ante.* Catherine in the subsequent Treaty of Jassy (1792) settled for a Black Sea fortress and its hinterland. Already the French Revolution and a new partition of Poland were capturing Europe's attention, and the Eastern Question passed unresolved to a later day.

(Da.H./Ed.)

ECONOMIC NATIONALISM AND MERCANTILISM

The growth of economic nationalism. The early modern age was a time of conscious state building. One after another, governments attempted to assert control over larger areas, and the new absolutisms tried to combine economic policies designed to achieve increasing national wealth with political programs aimed at increased power at home and abroad.

England was quick to achieve a high degree of unification, and this was confirmed as the Tudors strengthened their hold on Wales, Ireland, and the northern counties. Dynastic union with Scotland came with the Stuarts; and full legislative union in 1707. It cannot be claimed that the economic policies that were promulgated for this increasing area were always satisfactorily implemented, but a national policy was, at any rate, outlined in matters of tariffs, employment and poor relief, price and wage regulations, and the subordination of local authorities, municipalities, guilds, and companies to the will of the national government.

Union with Scotland

Elsewhere, progress was slower. During the reign of Henry IV in France, a significant practical contribution was made (in the shape of road and bridge building) to a concrete national economy. Under Colbert, principal economic minister to Louis XIV, something was done to unify weights and measures, to continue Maximilien de Béthune, duc de Sully's work of transport improvements, and to reduce divisive local tolls on roads and rivers. Even so, and in spite of French absolutism, France remained, in many respects, a large collection of provinces representing a great and varied wealth of natural resources but much divided by internal economic obstructions.

Even the Dutch republic, the model of economic modernity of the period, deferred to much local option in the shape of powerful cities and provinces. Real economic unity was achieved only by the predominance of the powerful and wealthy province of Holland. Both Spain and Italy faced insuperable geographical obstacles to any real economic unification, both countries being divided by mountain ranges. In general, the farther east the countries were, the more backward were their economies and societies. Germany in the 17th and 18th centuries still consisted of some 2,000 independent states of widely varying size, each with its duke, count, landgrave, bishop, or abbot among others. Russia, in spite of some rapid but exceptional projects of industrialization by Peter I the Great, was not modernized until after the Crimean War. Central and eastern Europe remained a society of lord and peasant (almost lord and slave), much of its business still conducted on the basis of barter or payment in kind.

Backwardness, however, was itself the stimulus to economic change. The examples of north Italy and the Netherlands—especially the Dutch republic—were constantly before the eyes of princes, bureaucrats, and even merchants struggling to cope with natural obstacles to transport, compounded by the remains of man-made obstacles surviving from years of fiscal impositions. In France and Germany, in particular, trade was constantly obstructed by internal tolls on roads, rivers, and canals. In 1685 it was calculated that out of a cargo of 60 logs, floated down from Saxony to Hamburg, only 6 remained after the customs officers at innumerable points had taken off their share as a tax contribution. Everywhere petty fiscalism obstructed the free flow of goods, and only at sea did cargoes move relatively freely (though even these were

Customs and tariffs

frequently subject to the attentions of privateers, pirates, and enemy warships).

The development of mercantilist theories. The 17th and 18th centuries nevertheless saw the rulers of these imperfectly integrated territories evolving a set of ideas with a remarkably high degree of common content. The theories were those that have come to be called mercantilist, though in practice they were as much the creation of ministers and administrators as of merchants, and their objectives were political and strategic as much as economic. In an age when war was almost continuous (the 17th century is said to have enjoyed only four years of peace), economic principles that rested on the assumption that wealth could be pursued in the abstract, without regard to the means to procure it, maintain it, and protect it by force, had little appeal to rulers. The mixture of politics and economics in mercantilism varied from state to state according to the resources, situation, and mode of government in force. In the Dutch republic the power of the state was most subordinated to the needs of trade. In England, France and, later, in Prussia and Russia, the economic initiative of the state was seen essentially as the indispensable other arm of the military and, often, aggressive intentions of the state toward its neighbours.

In spite of such differences of emphasis, however, mercantilist programs in most countries came to turn upon the concept known as the balance of trade. This concept was most clearly analyzed in the large volume of economic literature that appeared in England between the mid-16th and the mid-18th century. It may be discerned in the *Dialogue of the Common Weal* (1549–81) and is stated clearly by Thomas Mun in his *England's Treasure by Forraign Trade* (1664). The central theme is that, if the nation's exports exceed its imports, the nation will grow rich by reason of the influx of bullion that will result. But if the nation's imports exceed its exports, the nation will grow poor because of the loss of bullion that will necessarily follow. The next 150 years saw many refinements in the presentation of this simple proposition, as Mun's essay was discussed by scores of English writers and translated into French, German, Spanish, and Italian. But the basic concept of the "balance," supported by new bookkeeping practices and even by the notions of the new physics and mathematics, remained central to economic thought and policy down to the emergence of the classical economists and the ideas of the self-regulating economic mechanism introduced by David Hume and Adam Smith.

From this concept emerged, in turn, a series of economic measures designed to expand and adjust the national economy to meet the needs of the state and society. In its fully fashioned version, the mercantilist program aimed at encouraging the maximum influx of raw materials for industry by the removal of import duties and the imposition of export duties on local raw materials, wool being a frequent object of attention. Local manufactures, on the other hand, were to be encouraged by the imposition of protective tariffs against foreign competition and the grant of subsidies and export bounties to assist local manufacturers. An abundant supply of skilled labour should be provided by a large population, and, as the difficulties of calculating an exact monetary balance became fully appreciated, the main practical objective of mercantilists everywhere tended to become the employment of a larger and better trained labour force.

Export duties

This central core of ideas is common to the English mercantilists and to French writers such as Barthélemy de Laffemas and Antoine de Monchrétien in the time of Henry IV, the two most notable forerunners of the greatest French mercantilist, Colbert. They are repeated in the writings of the Austrian and German economists Johann Joachim Becher, Wilhelm Freiherr von Schröder, Philippe Wilhelm von Hörnigk, and the German cameralists, whose ideas powerfully influenced the ambitious economic programs of Frederick II the Great of Prussia. They appear constantly in the writings of the Spanish neomercantilists, Jerónimo Uztáriz and Bernardo de Ulloa, toward the middle of the 18th century.

In all these states successive attempts to turn mercantilist ideas into legislative form may be found between 1600 and the French Revolution in 1789. In France and in England such programs were most concentrated in the century between 1660 and 1760. The years from the Restoration to 1714 in England have been called the age of commercial and financial revolution. The emphasis there was on the regulation of trade (especially foreign), while industry was increasingly left to its own devices and largely freed from guild or government interference. The most important group of laws were those passed between 1660 and 1663, which developed in practical, enforceable form the ideas behind the Navigation Act of 1651—basically encouraging the growth of the mercantile marine by the compulsory channelling of trade between England and the outside world (especially British colonies) into English ships manned by English officers and crews. The use of foreign-built (mostly Dutch) ships was prohibited in order to boost English shipbuilding. Specified commodities of economic or strategic importance (sugar, cotton, timber, dyestuffs, naval stores) were enumerated so that their illegal carriage in foreign ships could be easily checked. At the same time, shipping was encouraged and agriculture boosted by the grant of bounties on exported corn (1673), and subsidies and protection were granted to a large range of industries including woollen textiles, linen, sailcloth, paper, metal manufactures, and many other products. Thus, a whole program of import substitution was launched and developed well before the Industrial Revolution. Between 1660 and 1700 exports and re-exports rose from rather more than £6,000,000 annually, while imports rose only from about £4,000,000 to £5,000,000 annually. This increase in export values continued, though at varying rates, until 1760, when they reached a figure of nearly £15,000,000.

English mercantilist legislation

The French program of Colbert was very different. Colbert was essentially the royal servant, ministering to royal needs, and his major task was to reform the tax system to improve the royal revenues. This he aimed to do by promoting manufactures through tariffs, privileges, and the import of skilled artisans, so that industry should be a source of national wealth and social contentment. It would be regulated through a universal guild system that would bring not only discipline to industry but also revenue to the crown and a reputation for high quality to French producers of luxury goods. Colbert was by no means the pacifist he has sometimes been thought, and he placed the whole of his regulated economy behind his master's aggressions in Europe, believing, as he did, that the amount of world trade was fixed and that France would only enlarge its share at the expense of the Dutch and English by force.

Financing of French wars

The results of mercantilist policies remain debatable. England certainly enjoyed a century of economic development. French industries likewise spread and, in some cases, seem to have benefitted, as did their English competitors, by the vigorous application of intelligence to their promotion. In an age when invention was limited and the price level stagnant or declining, it is impossible to deny that the mercantilists seem to have brought great stimulus to economic development. In Russia, the imitative mercantilism of Peter the Great has been seen by recent scholars as a remarkable example of state-promoted economic advance. In the Prussia of Frederick the Great, many attempts at unsuitable industrial development, such as silk manufacture, copied slavishly from French models, proved hopelessly unsuccessful, but the works of improvement to harbours, roads, and canals, the clearance of forests, the drainage of marshes, and the settlement of colonies of thousands of industrious Huguenots laid the basis for the industrial future of a united Germany.

It is no longer possible to write off mercantilism as an ignorant contradiction of self-evident economic truth as revealed by orthodox economic theory. Its political, military, and fiscal preoccupations were nevertheless dangerous, and, in an age when public opinion had only limited means to challenge the vested interests of bureaucracy, authority, and private monopoly, mercantilist institutions could easily become a hindrance to further economic progress. The French economy, in particular, was to suffer from the growing burden of ossified insti-

tutions, obsolete laws, and harmful privileges corruptly bought and sold.

Decline of Italy and the Dutch republic. The first half of the 18th century saw the completion of certain momentous developments. Internationally, trade moved away from north Italy and the Low Countries to centre more positively around the two foremost expanding nation-states, France and Britain. By 1700 there was no doubt that, commercially and industrially, Italy was in a state of relative decline. The great cloth-producing centres at Florence, Venice, and Como had all virtually ceased production, though silk manufacturing continued, often outside the old cities. Venice had finally ceased to be an entrepôt of world importance. In north Europe, the Dutch republic, heir to the Italian legacy, had lost its former dynamism. Essentially a middleman economy, it was now feeling the draft of competition from Britain and France. Merchants from these and other countries, who had formerly relied on the intermediary services of Amsterdam, were now trading directly with one another. Dutch foreign trade, while not positively collapsing, remained at about its mid-17th-century level. Agriculture even prospered, but the once-great industries of Leiden and Haarlem suffered badly. Although no longer the leaders of trade and manufactures, the Dutch continued to finance an important part of the needs of other states, especially for war. Dutch investors held an important stake in the rapidly growing English national debt, and there were few smaller governments that did not at one time or another turn to Amsterdam for loans.

Anglo-French competition

Italy and Holland may be regarded as examples of relative decline. Only Spain seems to present a picture of absolute decline from which there was to be no significant recovery even in the 19th century.

British economic and industrial growth. True expansion was evident only in Britain and France, and even there the early 18th century was punctuated by hesitations and losses. Everywhere population growth had slowed down. Cities frequently seemed to have reached the limits of expansion for the time being. London, for example, after a period of unprecedented population growth to over 500,000 people, had reached a point of overcrowding that threatened health and a point of general congestion on the Thames River that threatened profits. The new century brought about little expansion of its activities and with its stagnation went a levelling off of London demand for food and drink, necessities and luxuries. Breakneck expansion gave way to consolidation and more intensive development. These changes had their effects on agriculture, industry, and population throughout the Home Counties and East Anglia, and until 1750 landlords and tenants alike faced more years of depression than of prosperity.

Population growth

The total population of England and Wales may have grown by perhaps 1,000,000 to its 6,700,000 figure of 1760. It now seems probable that this continuing, though still modest, increase was due to a fall in deaths (33 per 1,000 in 1730 to 27 per 1,000 by the 1760s), whereas the birth rate rose from a low figure of 28 per 1,000 in 1710 to over 33 per 1,000 in the 1760s. In contrast to the deceleration of London, provincial cities had begun to grow. Norwich flourished on agricultural progress and its "new draperies," but the most significant vigour was to be found on the west coast, where Bristol, Liverpool, Whitehaven, and Glasgow in Scotland were all growing rapidly on the profits of the Atlantic trade—American and West Indian sugar, cotton, and tobacco and the provision of African slaves for the plantation economies of those areas. There was also a marked shift of manufactures, especially cheaper and lighter textiles such as linens, cottons, and fustians into the hinterlands of Lancashire and Scotland to meet the demand for exports to the new colonies. Ironware, hardware, nails, tin, glass, and small arms also filled the holds of Bristol ships sailing for West Africa, America, and the West Indies.

The western ports therefore accounted for a good proportion of the total increase in British shipping, which rose from about 320,000 tons in 1700 to just under 500,000 tons by the 1760s. Another important contributor to shipping demand continued to be the coal trade. Colliers taking northeastern coal from Newcastle to London had provided a major natural stimulus to English shipping and shipbuilding from 1550 to 1700. This cheap transport had, in turn, stimulated British coal production to a 14-fold increase by 1700. This expansion continued in the 18th century and was unparalleled in Europe; however, it took a new direction. Northeastern coal production, based on the London demand, grew less rapidly, while new coalfields, especially in the Midlands, ministered to the needs of an expanding pattern of output from a range of industries of widely varying type: sugar boiling, salt refining, brewing, and many small manufactures based on iron, such as tool making and lock, chain, nail, gun, and sword making. Abraham Darby discovered how to smelt iron ore with coke in 1709, but it took between another 50 and 100 years for the invention to penetrate widely throughout the iron industry.

Nevertheless, by the midcentury there were already signs that the growing markets at home and overseas were beginning to interact. As machines and engines called for more metal and fuel, coal output, in turn, responded, and this drew forth more demand for transport and labour. There was a notable increase in labour-saving mechanization everywhere, but, in spite of a few large-scale enterprises utilizing true factories, such as the Lombes's silk mill at Derby (1724), Wedgwood at Burslem (1759), and Boulton at Soho (1762), the new industries still grew by multiplying the number of relatively small units of production, employing new labour, and exploiting established technologies, such as dyeing, that were capable of improvement and expansion. All this new growth tended to provoke manufacturers and traders to challenge the older ideas of state paternalism and company monopoly, which now seemed to be a brake on output and profit rather than a help to their efforts.

Mechanization

Expansion of French trade and industry. In spite of the pattern of technical conservatism that state and guild control imposed on the French economy during the post-Colbert period, France also shared in the general expansion of trade and industry in the 18th century. From Labrador down to the Caribbean, the French participated in the trans-Atlantic trades, importing fish, furs, tobacco, sugar, coffee, indigo, and cotton. The French ports of Saint-Malo, Nantes, and, above all, Bordeaux rose on the new trades. Even Marseilles, though centred on the Levant and Mediterranean trade, benefitted from the re-export trade in West Indian sugar and coffee.

Certainly in the first half of the 18th century, the output of French cotton and iron was also greater than that of the rival industries of Britain. Some French industries were hindered by the rigid, monopolistic, secretive, and legalistic guild system that cramped enterprise and made it impossible to follow fashion. But elsewhere, French industry, like its English rival, sought expansion through rural development and the putting-out system. It enjoyed the great advantage of a large domestic market (three or four times the population of Britain and more than ten times that of the Dutch republic) and rich natural resources. Probably the expansion of the French export trade was even greater than that of the English, before the 1789 Revolution. French capacity for invention, especially in industrial chemistry and such ancillary processes as textile printing, was truly remarkable. Yet, in certain respects, French development may be seen in retrospect to have had less potential than its English rival. Shipping and shipbuilding were less closely geared to industry, and the same was true of the relationship between coal output and coal-using industries. The French pattern of manufacture was therefore less adaptable to the growth of mass demand. The custom known as *dérogation,* which penalized aristocrats who participated in trade or industry below the level of the great state colonial trading companies, may well have deprived the French economy of valuable financial aid and initiative of the kind that came from the shrewder members of the British nobility, such as that of Francis Egerton, 3rd duke of Bridgwater, the great canal promoter, or of the coal-owning Lowthers.

Anglo-French rivalry. Two other related weaknesses may be discerned in the French economic structure. En-

gland between 1688 and 1750 had developed a surprisingly well-balanced system of public finance, based on a well-managed public debt, an evolving system of taxation (both direct and indirect), the Bank of England, and the great trading companies, together with a widespread system of private banking. French financial institutions, on the other hand, did not succeed to anything like the same extent. Indeed, some contemporary observers saw in the weakness of French public finance and its inability to raise the great masses of capital represented by the growth in the English national debt (£12,000,000 to £132,000,000 in 1763) a major reason for British military and economic superiority that was to emerge decisively at the Treaty of Paris, which brought an end to the Seven Years' War in 1763. Militarily, France suffered in its overseas possessions—in the West Indies, North America, and India—by reason of this financial weakness, and military defeat was certainly reflected in overseas trade, not least in India, where sea power, especially, gave England the whiphand by 1761.

The national debt

The observer of these events in historic retrospect must nevertheless be careful. By 1750 the future of world power and of European control of trade both within Europe and in the colonial world beyond was still in suspense between the two great powers, France and Britain. It is not easy to determine how much of Britain's success in overtaking its rivals was inherent in the situation in 1750 and how much was enveloped in the events and developments that took place subsequently.

Modest though the advance of these two and a half centuries seems compared with what was to follow, it was by contemporary standards truly momentous. Europe's population very nearly doubled, to reach 140,000,000 by 1750. The growth in the area and volume of international trade and shipping was without precedent. The roles of capitalists and capital in the changing industrial system were as revolutionary as those of entrepreneurs and factories in the next phase of development. Finally, the men of the 17th century had begun to grasp how scientific method could be applied not only to the problems of agriculture and industry but also to those of society itself, a realization pregnant with possibilities for the future.

(C.H.Wi.)

THE ENLIGHTENMENT

The Enlightenment was a movement of thought and belief concerned with the interrelated concepts of God, reason, nature, and man that claimed wide assent among European intellectuals in the 17th and 18th centuries. Although diverse in emphases and interests, the Enlightenment attacked the established ways of European life and, in its conviction that right reason could discover useful knowledge, aspired to the conquest of man's happiness through freedom.

Ancestral roots. The roots of the Enlightenment date from the days when Greek philosophers discovered a regularity in nature and concluded that its governing principle was the reasoning mind and when, under the promptings of Socrates, they turned to consider man and ascribed a high value to his intellectual powers. Individual philosophers and their schools raised Greece to a brilliant eminence, but its glory faded and Rome emerged as the heir. The Romans saved much, adapted much, and made contributions of their own; but the Greco-Roman culture finally lost its vitality, and an unheralded tide of Eastern religions emerged, preaching doctrines of personal salvation. Christianity won the competition for converts, however, and before the end of the 4th century it became the exclusive religion of the empire. Christianity developed from an environment totally alien to that of the Greeks and Romans, and its purposes were so contrary to theirs that many fervent Christians wanted to repudiate pagan writings altogether. But the overwhelming majority of the converts were Greeks and Romans, for whom a radical break with their past was unthinkable; pagan literature was the only available source for philosophical justifications and for educational materials. And so Christianity had to make a cultural accommodation, but it tried to suppress the most immoral aspects of paganism and to bend the rest to the special purposes of the faith. The greatest achievement was that in the 13th century when Thomas Aquinas took up newly found manuscripts of Aristotle and made them into the very fundament of Christian philosophy.

Again, intellectual history moved from triumphs to reverses. More scholastic disputation badly injured Thomas' dominion; the papacy was moved to Avignon in 1309 (the discreditable "Babylonian Captivity"), and, worse still, expanding secular concerns in the 14th century shook the old order, nowhere more disruptingly than in Italy. Symptomatic of this change of atmosphere was Petrarch's newborn zest for classical letters and the response to it on the part of countless enthusiasts who roamed far and wide in search of forgotten manuscripts. Several revivals of classical letters during the Middle Ages had rescued many works from destruction, but the old religious purposes were not that of Petrarch and his friends, who wanted the ancient authors stripped of the distortions wrought upon them through the centuries of Christian editing and interpretation. These new men, the Humanists, sought to recapture the spirit of Greece and Rome, which had been expressed with such wisdom and beauty; and they produced a magnificent body of criticism and letters, historiography and science that made a glorious epoch in European history.

Petrarch's role in the revival of classical learning

But after about a century, intellectual life in Italy suffered ruinously from political turmoil, and Humanism went over the Alps to find congenial homes in Germany and elsewhere in western Europe. There the central purpose of Humanism was to serve religion, and this was especially true for Erasmus, who wanted to perfect a philosophy of Christ. Erasmus saw many things to reform, but when, in 1517, Martin Luther raised the standard of revolt, Erasmus drew back in horror.

The turbulent Luther unleashed more furies than he could have anticipated. A century later the old ecumenical church of the West was broken. The Protestants were appealing to the authority of the Scriptures, and the Catholics were invoking the authority of the organic church; the Council of Trent (1545–63) closed the door on useful dialogue between them. Meanwhile, the course of the Reformation had thrown up problems of its own, and debates over would-be solutions had scattered cross-purposes and confusion. By the 17th century Europe had no community of thought. The Renaissance had spread the diverse stream of Greek philosophizing, and Platonism, Stoicism, and Epicureanism competed for a hearing. Catholics and Protestants discharged their guns across an unbridgeable chasm; newer political questions raised up a veritable Pandora's box of conflicting opinions.

The decision of the Renaissance had been to keep the synthesis between classical letters and the Christian faith, but later developments had injured that hope. Inescapably, the need for an intelligible and accepted medium of communication sent men back to classical points of view, points of view not necessarily in conflict with Christian judgment but at the same time independent of it. One notable appeal to the minds of all Europeans was couched in terms of revived Stoicism. It was made by Hugo Grotius, a noble and scholarly Dutch jurist. In his *De Jure Belli ac Pacis* (1625) he invoked natural law as the rule governing international relations. For 2,000 years no one had denied the existence of natural law, but there was a fresh significance in the emphasis Grotius placed on its primacy and on its independence of Christian theology. This law, said the devoutly Christian Grotius, was so immutable that God himself could not change it.

Grotius' appeal to natural law

The scientific revolution. Another development was working out a new frame of reference—the beginning of modern science, a discipline that soon lost patience with religious quibbling and, no less, with the absurdities of Greek and Roman science.

Bacon and Descartes. The great prophet of this newest learning, the man who stirred the imagination of the next generations by painting the enrichment of human life through new discoveries, was Francis Bacon. Suspicious of the mathematics that had been making great technical strides, he proposed an all-comprehending method of induction. In contrast, Descartes drove from his mind everything except simple ideas, whose truth was undeni-

able, and he proceeded by rigorous deduction to work out his whole philosophic system. Despite the gulf between these two methods, Bacon and Descartes were allied in basic purpose and in influence. Both were at war with what they considered a harmful past; both dreamed of improvement in man's lot as the duty of science; both separated science from theology. Both, too, saw the need for practical experiment, and neither had a useful notion of how to go about it.

Galileo. The honour of fundamental pioneering belongs to Galileo Galilei, a man of genius, piety, and truculence. Taking up the old figure of speech that God had given man two books, Holy Writ and nature, he insisted emphatically tht the language of the book of nature was mathematics. For his purposes, Galileo limited his world of nature to properties capable of mathematical analysis, such as extension and motion, and therefore capable of being described in terms of precise natural laws. His method of research he elegantly demonstrated in working out the law of falling bodies by mathematical analysis of hypotheses and checking by experiment.

Such achievements should have been renown enough for any man, but Galileo had yet another triumph. As a starting point for it he had the astronomical work of Copernicus, Tycho Brahe, and Kepler. Their studies had done mortal damage to the Ptolemaic notion of a universe rotating around the Earth, but these great explorers had left behind them not only a storm of emotional resistance but also a number of crucial unanswered questions. Report had come to Galileo of a new invention called a telescope, and he made one for himself. In a few nights of peering into the heavens he saw a system of structure and motion that proved the Copernican heliocentric explanation to all but the most hidebound defenders of the old order. With the discredit of the Ptolemaic astronomy, the whole of Aristotelian natural philosophy became untenable, and Galileo in 1632 recklessly set down his views in *Dialogo sopra i due massimi sistemi del mondo, tolemaico e copernicano.* He dedicated his book to the Pope, but that act of courtesy availed him little. He was called before the Inquisition to hear a verdict already prepared; but the broader significance of the trial was that Galileo's new scientific way of thinking had gone so far from inherited modes that no reasoned interchange was possible, and the church based its condemnation on fear.

Newton. Another difficult problem remained: the need to construct a general system of mechanics that could account for the motions of the stars in terms of observed behaviour of matter on Earth. A number of scientists worked at it, but it fell to Isaac Newton to win the honour of success. In 1687 the Royal Society, under the authorization of its secretary Samuel Pepys, published Newton's *Philosophiae Naturalis Principia Mathematica.* In this work, with prodigious mathematical reasoning, Newton presented the law of universal gravitation.

The effect of Newton's discovery of the law of gravitation

Newton modestly confessed that he did not know what gravitation was, and when his calculations indicated some irregularities that seemed to require the direct intervention of God to set them right, he was honest enough to say so. The German scientist and philosopher Gottfried Wilhelm Leibniz, who had quarrelled with him over priority in the invention of the calculus, attacked Newton harshly for picturing God as a clumsy watchmaker, but few had patience with such small matters. Newton had explained the universe; he had vindicated the rationality of nature; and he was, said the zealous Voltaire, the greatest man who had ever lived. Behind this enthusiasm for Newton and for natural philosophy there lay a serious question. The mounting tide of discoveries in physics pressed toward the autonomy and the moral neutrality of science, but the spirit of the time could not allow this kind of isolation for so wonderful an instrument of knowledge. The by-product of science had been the destruction of wide domains of medieval error; perhaps its methods, its experimental reasoning, could be used for all of man's problems. Indeed, Newton himself had suggested the possibility.

Enlightened religion. The Christian faith showed a powerful tenacity when confronted by this avalanche of new facts and new conclusions, but it could not fail to be acutely sensitive to the changing intellectual fashions. By the early 18th century, radical ideas hitherto submerged came into the open and precipitated a debate fraught with great consequences. Hitherto, the rationalists had had no base for comparison except their knowledge of the ancient world. Latterly, they had become acquainted (in considerable measure through Jesuit missionaries) with pious Egyptians and Siamese and with Chinese, who, they said, had no real religion at all but were still men of exemplary virtue. More recently still, another stock character came upon the European stage: the noble savage, truly a child of innocence who worshipped his god in simple faith and obeyed the precepts of Mother Nature. And, finally, this growing breed of scientists was steadily grinding out disturbing implications. Good Christians though most of them were, they showed that Christianity's views of nature could not stand up under the tests of experimental reason.

The rise of Deism. Such considerations led dissatisfied men to assume that, beneath the world's religious diversities, there was a common body of beliefs, a religion planted by nature in all men everywhere. In 1624, while Grotius was preparing his case for natural law in international relations, Lord Herbert of Cherbury published his *De Veritate,* a definition of natural religion. He said that certain beliefs are so manifest that all men of reason accept them: the existence of one God who dispenses rewards and punishments, and the obligation resting on man to worship him in repentant piety and virtue. These articles sufficed for a religious life in the present world and for salvation in the hereafter. Deism, as this doctrine came to be called, did not always radiate Herbert's lofty Stoic tone, but his statement was so representative that it persisted until Thomas Paine, in his *Age of Reason* (1794–96), wrote the swan song of the movement.

The Deists' views

The Deists maintained the old synthesis between God, reason, and nature, but, reasoning from a mathematically ordered nature to its architect, they allowed the architect only the qualities manifest in his handiwork. Grand metaphysical systems fell under the ban as mere extravagant nonsense; philosophizing, said the dominant voice, should follow the method of the natural sciences. Such convictions led inescapably to relentless warfare on those beliefs that were unique to Christianity. The Holy Writ of the Christians suffered denunciation as groundless historical confusion; but the Old Testament came in for special attention—it was scornfully dismissed as a repository of the crimes and obscenities in which God connived with the so-called chosen people. With equal vigour the Deists ridiculed biblical prophecy out of court; the notion that God would make revelations in such a book, or in later manifestations, was utterly offensive to the rational mind. Miracles, therefore, were unthinkable. It would do violence to reason to suppose that God would, or could, set aside the laws that gave expression to his own eternal being.

Deism was never an organized movement. Its first home was among English intellectuals. Around 1700 a series of books appeared that hit ruthlessly at what they held to be the misconceptions and offenses of orthodox Christianity and argued for a depersonalized God, whom anyone could discern for himself. Some of the authors offended good taste and the law by their blasphemies; some wrote with calculated prudence. But during a half century the central core of their argument remained essentially unchanged: Christianity was an indefensible burden on the back of natural religion. But then English Deism went quietly into decline, in part because public disputation was reduced to tedious repetition and in part because the Anglican Church was too preoccupied with political advantages and social dominance to wish to disturb its fat slumber by theological contentiousness. England in 1688–89, with its Glorious Revolution, had settled its wracking political issue, and it was time to relax. The clergy preached coolly on the prudential morality of the Christian life, while Joseph Addison taught his readers the affable urbanity appropriate to a civilized society in an age of reason. Deism moved its headquarters to the Continent.

France had its own native roots of Deism, which grew vigorously on the disillusionments and discontents of the reign of Louis XIV. Voltaire was already committed to

an anti-Christian Rationalism before he sought a haven in England in 1726. On his return after two years of exile he wrote a sparkling report, *Lettres philosophiques,* which reaffirmed his commitment to the Deist view of religion. Through his remaining years, his faith remained inflexible. Pressed by more radical opinions, he held steadfastly, in the end somewhat frenziedly, to the God whose evidence he saw in nature and whose existence he deemed necessary to preserve the social order. His one change of emphasis was in the growing vehemence of his anti-clericalism. In the 1760s he completed his *Dictionnaire philosophique,* a witty but earnest compendium of malice toward Christianity, and at the same time, poltroon though he frequently was, he entered the public lists to bring justice to several persons he considered to be victims of ecclesiastical murder. By then he had adopted in rage the motto *Écrasez l'infâme!:* "Every sensible man, every honourable man, must hold the Christian sect in horror."

No one was more temperamentally removed from Voltaire than Jean-Jacques Rousseau, but when he included a profession of Deist faith in his novel *Émile,* Voltaire sent him an enthusiastic compliment. Rousseau's profession was a manifesto for reason as the only reliable test for religious truth and equally an attack on all doctrines and clerical institutions that stood between autonomous man and God. Rousseau characteristically made a place for heightened emotion in such passages as the invitation to go to the sublime book of nature and learn there to worship the Creator; but the overtone of anti-intellectualism did not obscure the rational doctrine of Deism. Throughout France opposition to the church

The spread of Deism

spread, drawing strength from the debilities of the *ancien régime,* and the Deist faith went on to the culturally less well prepared soil of Germany and to the British colonies in America. The movement kept up its momentum far into the second half of the 18th century, but it then paid the price of its lack of organization and the weakness of its emotional impact on uncritical minds. Without a sacred text, without a compelling leader, Deism was always a matter of individual judgment and thereby vulnerable to centrifugal forces.

The Christian response

Naturally, when established societies felt the shock of Deist subversion, they resorted to the time-honoured methods of defense. First came Christianity's dogmatic assertion of truth from pulpit and press (*e.g.,* in 1770 French publishers brought out 90 books in support of the Catholic faith). Behind such gentle measures, however, there were, in considerable diversity, grimmer ways of suppressing dissent. Grimmest was the vengeance of the Inquisition, but hardly less so was the possibility of civil judgments, including the death penalty. And then came the *Index Librorum Prohibitorum* (*Index of Forbidden Books*), ecclesiastical and state censorship and licenses to print, and book burnings. Much skilled and harsh determination went into devising means to stifle corrupting thought. Many individuals suffered many punishments, but heresy was a hydra-headed monster, thriving on persecution and on the stupidities of its enemies.

In Vienna in 1772 the political authorities suddenly realized that they had to put their own *Catalogue of Forbidden Books* in that self-same *Catalogue,* because the impure of heart were finding it a useful guide to spirited reading. Administrative inefficiency and, notably in France, competing authorities helped to dent the cutting edge of tyranny, but the greatest aid for those dealing in illicit materials was their own ingenuity in adapting the arts of the smuggler and the forger into their affairs. Books sold under the table, books without authors, books with false places of publications—all were commodities that did well. In these deceptions Voltaire was by no means alone, but he was a superior master. His brilliant style was more clearly his own than was an acknowledged signature, but in edition after edition he denied writing the *Dictionnaire philosophique,* and no policeman challenged his audacity.

Such cat-and-mouse antics were not the whole story of this confrontation. The real trouble was more subtle. Christianity was rooted in both reason and revelation, and, according to the Fathers and the doctors of the church, these sources were not in conflict; revelation simply had the higher truths. Now came Deism, professing to have the full credentials of reason and demanding that the Christians show how they could, in reason, defend revelation. Again the first battleground was England. Increasingly, supporters of the Christian faith found themselves driven to admit the validity of natural religion and then to search out reasonable grounds for adding revelation. John Tillotson, future archbishop of Canterbury, spoke for many when he stated that Christianity was needed to complete the findings of natural religion: its service was to test the reasoning process and to give a supplementary motive for a moral life. Tillotson conceded much, but his contemporary and fellow Christian John Locke went even further when he accepted reason as the judge of things above reason. Bishop Butler's *Analogy of Religion Natural and Revealed* (1736) offered a fresh line of Christian defense. It became a famous book, but, despite the kindly Bishop's Christian dedication, it did not demolish many Deists. In the pinch, he did not venture into dogmatic assertions based on the usual Christian sources; he took his stand on probability, and he reminded his readers that there were things in natural religion and in nature just as difficult to believe as was orthodox Christian doctrine: after all, Newton could not explain gravitation. Comparable gestures toward a friendly view of the rational standard appeared in Holland, in Germany with a group of prominent Lutheran pastors, and in France, where the Jesuits had a high skill in reasoned debate until their expulsion in 1762. The consequence was that, on the Continent as in England, the theological content of practical, daily religion declined significantly. The popular preachers discoursed in muted tones of reasonableness on the beauties of Christian virtue.

While Deists and Christians were exchanging blows, they tended to agree that God in his benevolence had designed the Earth for man's happiness; "the best of all possible worlds," Leibniz said. Another said that the sun rises and shines for men, and still others pointed to the evidence of the Creator's forethought—the density of water to facilitate navigation; the shape of the melon to make it easy to cut. Even Voltaire, having no conception of change in the universe, surveyed the mountains and saw in them the wisdom and the benevolence of God.

Atheism. Both Deists and Christians heard with anxiety that their happy assumptions were threatened by the upsurge of atheism. Materialism was an ancient philosophy. Epicurus had propounded it, and his Roman disciple Lucretius had put it into his poem *De rerum natura,* which, still later, had excited the Renaissance enthusiasts for classical letters. Throughout the whole of its history, Epicureanism had been a scandal for the mass of literate people, but a notable moment came in 1563, when a French Humanist published a new edition of Lucretius and urged his readers to forget the absurd ideas and concentrate on the poetry. The plea helped a bit, but a far more significant work of salvage was done in the 17th century by Pierre Gassendi, a Catholic priest of blameless repute. Boldly, Gassendi said that Epicurus had not been an advocate of vice but had cultivated the virtues of honesty and inner peace. His Materialism, Gassendi went on, was erroneous; but he shrewdly suggested that it could be made palatable by postulating Divine Providence behind Epicurus' atoms and voids. Descartes would not admit a void in his universe, but he made a contribution, however unwillingly, by his dualism of spirit and matter, which made all animals, including physical man, into purely mechanical structures. With the growth of science, the 18th century took kindly to Gassendi's revival of atomistic philosophy and to the Cartesian mechanistic view of nature. In England David Hartley rewrote Lockean psychology and explained mental life in terms of physical vibrations building complex patterns by the equally mechanical processes of association. In France another physician, La Mettrie, brazenly spelled out the title of his book, *L'Homme-machine* (1747), and stated that man, like everything else, was reducible to matter in motion. "We are no more committing a crime," he wrote, "when we obey our primitive instincts than the Nile is committing a crime with its floods, or the sea with its ravages." Voltaire would not entertain what was to him

The rise of Materialism

an offensive notion, but the Encyclopaedist Denis Diderot was tempted and succumbed. The most powerful atheist influence was that of Baron d'Holbach, a wealthy German who had settled in Paris to wage war against God. He commissioned the publication of a whole library of books in addition to those he wrote himself; his only criterion was that a book should serve the cause. In his eyes, religion was the main source of degradation. Rumours circulated about his dinner parties, and ordinary folk felt uneasy when they passed his house, but d'Holbach went further in his anger against Christianity than many men were willing to go.

Skepticism. One further, and more powerful, threat to both Deism and Christianity was skepticism, also a philosophic doctrine that had ancient roots. Because of the lasting popularity of Montaigne's essays of the 16th century, one must see in the meditations of this gentle devotee of ancient letters a fresh source for modern doubt. As he mused on human folly and fanaticism, he repeatedly threw in his personal conclusion, "What do I know?" It was a diffident question, but over several generations it prodded many minds.

The influence of Pierre Bayle

A century later, subdued reflection gave way to a noisy uproar. The cause was Pierre Bayle, whose course from Protestantism to Catholicism and back to Protestantism sent him into a life of exile and of hostility to all religions. His instrument of vengeance was his *Dictionnaire historique et critique,* the first edition of which appeared in 1697. In form it was a collection of biographical sketches with comments, but the individuals chosen and the amount of space given to each reflected Bayle's deadly purpose—to demolish the vices of religion. With learning and wit and with a titillating spice of salaciousness, Bayle drove his sword at every exposed surface of the erstwhile eternal verities. Perhaps nowhere did he hit harder than at the Old Testament when he asked how it was that, when Holy Writ chronicles the morally offensive record of King David, it goes on to call him a man after God's heart.

Bayle's *Dictionnaire* had an extraordinary success. It was soon expanded in new editions, translated into English and German, and boiled down into abridgments and digests. The rollcall of those who read it was a veritable *Who's Who* of the 18th century. Frederick II of Prussia commissioned two abridgments; Benjamin Franklin and Thomas Jefferson recommended it to their friends; Voltaire drew inspiration from it for his own *Dictionnaire philosophique;* Diderot wrote a lyrical article on Bayle for the *Encyclopédie* and was enraged when the rascally publisher suppressed the main substance. No one, however, was more persuaded than the Scottish former Presbyterian David Hume.

At 28 Hume published the first two volumes of *A Treatise of Human Nature.* The subtitle showed the cast of a mind that never changed: *an Attempt to introduce the experimental Method of Reasoning into moral subjects.* From this first, and unsuccessful, book to his posthumous *Dialogues concerning Natural Religion* (1779), Hume, with amiable reasonableness, dissected all certainty from human affairs. Down went the concept of cause and effect, an idea seemingly essential to all rational thinking; down went the celebrated powers of human reason. But the realm of belief suffered most notoriously in this raging storm of doubt. The miracles of so-called revealed religion he subjected to the test of experience, and they dissolved beyond rescue except for the most willful believers. Even natural religion could not stand up to the test of fact; there was no universal consensus supporting it. According to Hume, primitive men, in their fear, invented gods, many gods, to gain protection, and it was only a long time later that monotheism emerged from ignoble calculations. The great trouble, said Hume, was that men assumed that the Supreme Being gave his qualities to them as an endowment of the creation, whereas in reality men ascribed to God their own values. Here, then, was a repugnant philosophy; nothing was left but mere uncertainty and probability at best. The only saving grace, Hume puckishly admitted, was that one could not be a skeptic all the time.

Christian defenses. In the eyes of orthodox believers in Christ, the new trends were all manifestations of a perverse spirit that incited defiance as well as consternation. From Christians little moved to dispute came an assertion that their religion did not need to submit to these debasing tests of rationality. The mystics and others of unassailable conviction went their own ways, indifferent to foolish clamours. Still, it was not always easy. Blaise Pascal, mathematical genius and convert to the rigorous faith of the Jansenists of France, had felt agonizingly the tension between the old ways and the new. He had been convinced that human reason could not explain the ultimate philosophic questions, and he sought for moral certainty in Christian revelation. The universe was vast and men were insignificant, but they had souls and they were dependent on God. In his fear of the vastness of space, Pascal committed himself to a supreme act of faith: "The heart has its reasons which reason does not know." As a student of the law of probability derived from the gambling table, Pascal took refuge in his wager on the existence of God: "if I lost I would have lost little; if I had won I would have gained eternal life."

Chief of those who a century after Pascal would not accept the threatened demolition of the Christian faith was John Wesley, an Oxford student of the classics, early given to piety and a second-time convert through the Moravian Brethren (a German pietist sect). It was no novelty when Wesley said that reason could not produce the love of God, but it was strangely exhilarating when he told his spiritually hungry followers, "A true and living faith in Christ is inseparable from a sense of pardon from all past and freedom from all present sins." For 50 years Wesley preached his simple message of the love of God manifested in Christ, and the humble folk found consolation as Wesley travelled up and down England preaching an average of 1,000 sermons a year.

And so before the end of the 18th century something less than success had crowned the efforts of those earnest men who used emancipated reason and the methods of science to explain God and the ancient mysteries of the universe. The old faith had been buffeted, but Wesley's was only one of the first announcements of a religious trend toward the unleashed emotional flights of the Romantics.

The investigation of man. Meanwhile, the attempt to examine the nature of man in the light of new thought raised difficult problems. The old and continuing Christian view had a harsh consistency: man carried in him the stamp of Adam's sin until the hour of redemption through Christ's atonement arrived. The Renaissance concept of the dignity of man was an implied criticism of the old belief. But now, in the heady days of rationalistic deductions, it seemed axiomatic that God had implanted his reason and his goodness in man. In this view there was no place for original sin; were it otherwise, the creation would have been monstrously unjust, and that was unthinkable. But the difficulty about this happy analysis was to find men who were convincing demonstrations of the thesis.

The notion of the goodness of man

Locke's epistemology. Because a priori reasoning about man was not very profitable, it was in order to utilize Newton's new scientific method to study human nature. The pioneer was John Locke, and his vehicle was *An Essay Concerning Human Understanding* (1690). Locke's purpose was to determine the scope and limitation of human knowledge through an investigation of the workings of the mind. In the end he set a low limit to knowledge, but what made him the great teacher of the 18th century was his description of the mental process: all knowledge arises from experience, from sensations sorted and combined by reflection. There are no innate ideas; every child is born with a mind like a blank sheet of paper on which the impingement of the external world inscribes the data of life's accumulation of learning. Bishop Berkeley attacked Locke for assuming a world of substance as the source of sensations when that world was beyond the reach of knowledge; but aside from Berkeley's criticism, the psychology of the next century was essentially a gloss on Locke.

The Lockean investigation had demolished not only innate ideas but also innate depravity. Environment made the man, and the most obvious corollary was that the way to improve man was to improve the environment. It was a lesson that gave new dimension to human possibilities

and therefore a strong impetus to the zeal for humanitarian reform.

Ethical questions. Locke was primarily concerned with epistemology and gave little attention to the nonmental aspects of man. The language of the day spoke of the passions—the emotions—and their role became increasingly important as the study of human behaviour continued. The rationalist deduction held to man's inherent goodness by virtue of his reason: to know what was right was to do what was right. This statement hardly fitted the observable facts, and a qualification was added: when reason was not obscured, it would tell man what was right and he would act accordingly. But, again, this revision was short of full accuracy: it was established fact that a man who clearly knows the truth may not serve it. It was therefore in order to investigate more carefully the inner sources of moral action. Even exalted rationalists had to admit the existence of ethical promptings in the nonintellectual side of human nature (*i.e.,* in the passions), and two generations and more sought a precise explanation. Early in the 18th century, the 3rd Earl of Shaftesbury found it in an instinctive love for the good because of its inherent worth, its beauty, and its reflection of nature. Bishop Butler pointed to the conscience, the voice of God speaking to the human soul. Even David Hume turned from his skepticism long enough to ascribe to man an innate sentiment for humanity, and his friend Adam Smith, the calculating economist, gained his first celebrity in 1759 from a book that asserted that sympathy was a guiding emotion in human conduct. Across the channel, Voltaire, that alleged prince of cynics, found in man a natural benevolence never seen in the beasts, and the radical Morelly wrote in his *Code de la nature* (1775), "In the natural order the idea of active or passive benevolence precedes every other idea, even that of Divinity."

Running through these comforting reassurances was perforce a question of the relation between this ethical faculty, rooted in the emotional impulses, and reason. Shaftesbury was persuaded that rationality and moral emotion could never be at odds because both were the gifts of nature. Butler contended that conscience had its own power of reasoning. In the second half of the century the common voice began to falter. Adam Smith could not say that sympathy and reason would invariably counsel the same moral decision, and Hume, returned to his most destructive vein, wrote in his *Treatise,* "We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be, the slave of the passions, and can never pretend to any other office than to serve and obey them." Many of Rousseau's passages implied, however unphilosophically from Hume's point of view, an open conflict between reason and the passions. Under the inspiration of Rousseau's idealized man of virtue, novelists were soon playing variations on the theme that goodness was merely a matter of opening the heart's emotional spigots.

Conflict between reason and the passions

Meanwhile, the discussion of ethics had addressed itself to an even more important theme. Christianity had always made its basic appeal to man's desire for the happiness of heaven and his fear of the torments of hell. It was easy to generalize this human predilection and say simply that the dynamic force in life was the desire to secure pleasure and to avoid pain. Christian John Locke agreed: "Pleasure and pain and that which causes them, good and evil, are the hinges on which our passions turn." With a boldness that threatened to become unseemly, numerous writers, notably the exuberant Diderot, sang praises of the passions and in so doing obliquely discounted reason.

It was necessary to think of the ethical implications. Few were ready to make the pursuit of pleasure a moral code in itself. The extravagant determinists could argue that man's decisions and acts were simply the consequences of matter in motion and that there could thus be no question of morality in human life; but most reflections on ethics, despite Hume, assigned to reason some kind of brake on personal misbehaviour. Others contended that the experience of living with people taught men the need they had of each other and the ensuing necessity for reciprocal obligations. In addition, as a justification to some and a reassurance to others, there circulated an assumption of a mysterious harmony between self-interest and the common good. Montesquieu wrote, "Everyone pursues the common good under the impression that he is following his own private advantage." The impudent Bernard de Mandeville expressed the same conviction in his *Fable of the Bees* and summarized the whole matter in his pithy aphorism: "Private vices, Publick benefits." Alexander Pope was more eloquent: "Thus God and Nature link'd the gen'ral frame/And bade Self-love and Social be the same."

The brave Newtonians had confidently set out to create a science of morals, but they had only created confusion. Another question arose: if there was this anxiety about the capacity of the individual to find within himself a reassuring incentive to ethical conduct, was it possible to find a source external to the individual? For believing Christians the answer was simple: God through the Holy Scriptures and his providential governance of the world has given his commands, and it was man's duty to obey them. It was a far more difficult task for the would-be scientific thinkers, who had broken relations with the Christian God. Thomas Hobbes in the mid-17th century had left a suggestion. In a bit of hypothetical history he told of people who in the state of nature were, each and every one, under the sovereignty of pleasure and pain. The consequence was war of all against all, and life was "solitary, poor, nasty, brutish, and short." To escape from this deadly plight, these people set up a state and consigned to it an all-comprehensive power to rule. The compact created obligation. Hobbes's successors faced the same problem. To restrain moral anarchy it was imperative to design a code based on the welfare of the community. The expression "the greatest happiness for the greatest number" had first appeared in the 1730s, and Claude-Adrien Helvétius made an ominous suggestion. In his book *De l'esprit* (1758), burned by the parlement of Paris and anathematized in Rome, Helvétius explained how to take this self-centred creature man and make out of him a social being: "It is solely through good laws that one can form virtuous men. Thus the whole art of the legislator consists of forcing men, by the sentiment of self-love, to be always just to one another." Out of such materials Jeremy Bentham drew his blueprint for a laissez-faire democracy.

Humanitarianism. It was perhaps less important that these 18th-century thinkers could not fashion an agreed science of morals than that so many people were deeply concerned. The Deist search for natural religion had led to wide-ranging investigations of peoples all over the Earth, with investigations centred not on manifest differences in colour and customs but on what people had in common. And they found much. "It is universally acknowledged," Hume wrote in 1748, "that there is a great uniformity among the acts of men, in all nations and ages, and that human nature remains still the same in its principles and operations. History informs us of nothing new or strange in this particular. Its chief use is only to discover the constant and universal principles of human nature." The sum totality of all men made up humanity, and humanity, by some subtlety, inspired loyalty and compassion. Under this influence, vaunted reason was turning itself unobtrusively into reasonableness. When the concept of natural law was applied to social problems it underwent a significant modification. It no longer stated what was, but what ought to be; and since in so many respects the human condition fell far short of the ought, the Enlightenment became critical, reformist, and eventually revolutionary.

Universal principles of human nature

Locke's demonstration that environment made men provided the dynamic inspiration for a vast attack on prevailing abuses. In the presence of disinterested compassion, those laboured analyses of man chained forever to his pleasures and pains seemed sterile and irrelevant. The long, ceaseless campaigns waged by the intellectuals for freedom of religion and speech might appear to a cynic to be breast-beating exercises to secure their own privileges and rights, but certainly the cynic's yardstick could not measure a countless list of self-jeopardizing acts by men working to increase individual freedom for the simple reason that individual freedom was good. Out of that same

simple reasoning came an attack on the inhumanity of Europe's brutal administration of brutal law.

One blind spot, as later generations would call it, was the Enlightenment's little thought for the misfortunes of the lesser rungs of the European social ladder. The poor Europe had had with it always. Sympathy leaped more readily over the horizon to embrace people of dark hue whose status touched the European conscience. Atheist and Christian equally agreed that slavery was morally wrong, and they made common cause against it.

The religious radicals and their Christian enemies also shared an insistent sense of all men linked together in universal brotherhood. Standing in the way of the ideal were many grim realities, but the most baffling was the sovereign state, which demanded the obedience and, far too often, the blood of its people. Despite early warnings of a ferment of nationalism, men of many divergent persuasions toward other issues sounded their defiant calls to world citizenship. Invoking the spirit of Grotius, the Swiss lawyer Vattel pleaded for a society of states living together in peace under the morally binding prescriptions of natural law. Other men who loved peace recalled a proposal attributed to Henry IV of France to establish a formal international confederation that would make war in Europe impossible; the Abbé de Saint-Pierre revived this project in 1713 and worked at his cause for 30 years without gaining many converts. Rousseau published a rather lukewarm restatement, and there the peace campaign rested until 1795, when Immanuel Kant spoke the final word for the Enlightenment in "Zum ewigen Frieden." But Kant's eloquence was to no purpose, for at the time Europe was already two years into a war that would not end until 1815.

Social thinking. The application of the Enlightenment's rationalistic and scientific tools to social fields led to no settled agreement, either as to method or to conclusions.

Political thought. The Reformation had thrown up doctrines ranging from primitive communism to the divine right of kings. Bishop Bossuet restated with the best French eloquence of the 17th century the special relation between God and Louis XIV, but the ancient principle of the natural law went on talking about the state of nature, the social contract, and natural rights. The weight of these theories were, by their very nature, on the side of radicalism. Yet a curious thing happened—the first notable statement of political philosophy in the new age of mathematical rationalism was a secularized theory of absolutism. The author was Thomas Hobbes, who became a rigorous reasoner when, in middle age, he chanced upon Euclid's *Elements.* Hobbes started with natural law but, as indicated above, he admitted no "ought to be." The law imposed a sleepless search for pleasure and the passionate desire to escape pain. The people caught in the desperate consequences of this rule of life made a social compact whereby all resigned their individual powers into the hands of a sovereign outside the compact, whose authority for all practical purposes was unlimited. Hobbes had many acute insights into human affairs, but Charles II, whom he was trying to support in the distress of his exile in France, was so scandalized that Hobbes had to flee the royal presence and take refuge with Oliver Cromwell.

The social compact

Half a century later John Locke accepted Hobbes's strong sense of individualism and his conception of the state as a mere mechanical convenience rather than a divine institution, but in all other respects his arguments went back to pre-Hobbesian terms. The state of nature, said Locke in the second of his *Two Treatises of Government,* was a social regime of men of good will; in it, natural law had endowed men with rights, principal among which were life, liberty, and property. Certain confusions arose when every man defined for himself and tried to protect his own rights, and it was minor difficulties and not desperation that inspired the formation of a political community specifically and exclusively for the protection of individual rights. Government was derived from the consent of the people, and it functioned by majority rule. Locke's purpose had been to compose a challenge to Charles II's brother, who was heir apparent; but the *Treatises* were not published until 1690, and by that time the Glorious Revolution had closed a troubled period and Locke's revolutionary tract was appropriated as a defense of the Revolution. Thereafter, England settled down for a long time to enterprises requiring no new political philosophizing.

While England was relaxing into complacency, most of the Continent still remained too bound to a lethargic inheritance to think of formal political theory. A few Germans, a few men of the Low Countries, an occasional Italian rose sufficiently above their bleak routines to more systematic thought, but it is best, for summary purposes, to concentrate on the affairs of France. Long before the death of Louis XIV, times were out of joint and remained so, with few respites, to the end of the monarchy. There were paralyzing difficulties, however, in finding the causes of the disturbed temper of the 18th century and proposing suitable remedies. The French monarchy had done too well in that it had stifled all prospects of generally popular institutions and had made public discussion a hazardous occupation. The monarchy had done badly in that it had not rooted out the mentality and the vested interests and important agencies of feudalism. The monarchy had also done badly in that its close ties with the Catholic Church made every grievance against the church either obliquely or overtly a grievance against the royal structure of power. The consequences were disastrous for judicious thinking about politics.

All his life Voltaire unremittingly poured his energy into supporting the powers of the crown against the parlements, the medieval law courts that had become citadels of aristocratic resistance. Montesquieu wanted a greater role in government for the nobility of which he was a member; for that reason he held the British House of Lords in high esteem. Diderot, not by instinct a political animal, never understood the contest between king and nobles and always blew hot or cold as immediate circumstances prompted him. And yet for him, as for Voltaire, realism bade him acknowledge that all kings of France, even Louis XIV, had faults; but he never revealed doubts about the monarchy. He had courted Louis XV for preferments and had got them, and he took pride in the fact that Frederick II of Prussia had courted him. A new trend stretching across Europe from Spain to Russia gave plausibility to the thought that man's best hope for reforms lay in the vigorous sovereign. In the most unexpected capitals, rulers were reorganizing their governments, removing tax and other burdens on trade, and restraining the brutalities of the law. No enlightened Philosophe in France could complain until he realized that these so-called enlightened despots had another, and a repugnant, face turned toward war and conquest.

What was left for France to talk about, and to write about on condition of prudence, were the abuses of political power along with those of ecclesiastical power. Despite this dejecting atmosphere, in 1762 a new kind of political book appeared in Paris. It was called *Du contrat social* and its author was the difficult, emotionally unstable Jean-Jacques Rousseau, who had first won fame by a perfervid attack on the evil consequences of civilization. Rousseau proposed a society able to cultivate the individual's moral stature without injury to his freedom. It was a fresh effort at that knottiest problem of political philosophy. For a solution he returned to the old, and now somewhat derided, idea of a social contract. According to Rousseau, each individual ceded his rights and powers to a general will of which he was part—that is, an equal participant in the political life. As he stated the terms, they logically ruled out a distinction between rights and powers reserved to the individual and the authority and functions assigned to the community, that issue which was of paramount importance to Locke. By Rousseau's reckoning, the individual was as free after the contract as he had been in the state of nature; he had merely substituted for natural freedom a civic freedom by which he was elevated into a moral being. Some later readers have found in *Du contrat social* dangerous undertones of totalitarianism because of the equivocal place of the individual in relation to the general will, because of the axiom that the general will was right, and because of a prescription of a civic religion with penalties for defiance. Other latter-day students have

Rousseau's *Du contrat social*

rejected the imputation. It may be permissible to conclude that Rousseau, in writing ostensibly a book on political theory, did not really have his mind on politics. In one respect, nonetheless, he gained in clarity over John Locke. The *Two Treatises of Government* had referred to the powers of an undefined majority; *Du contrat social* wound up in an emotionally suffused claim for political democracy.

Both men had their triumphs. In America Thomas Jefferson incorporated strict Lockean doctrine in the Declaration of Independence. When shortly thereafter the French revolutionaries wrote their own declaration, they made a place for Rousseau's general will and sovereignty of the nation. On both sides of the Atlantic, however, there was fumbling over the fundamental question of popular government: who was to have the right to vote.

The mainstream of Enlightened thought had no quarrel with Locke's assertion that property was a natural right. Yet Rousseau, in agreement with Hume, was uneasy. Although he did not write down specific remedies, he was persuaded that political democracy could not work without a relative economic equality. In the seething atmosphere of the second half of the 18th century, a little band of dissenters took their criticisms far beyond Rousseau's doubts. In 1755 the obscure Morelly, in *Le Code de la nature,* attacked property as the mother of all crimes tormenting the world, and he made a proposal that was to have a long history: let every man contribute to the commonweal according to his abilities and receive according to his needs. Two decades later the Abbé de Mably, starting with equality as the law of nature, contended along with Morelly—and, for that matter, with the early Rousseau—that the introduction of property had destroyed the golden age of man. Soon, François-Noël Babeuf, civil servant turned radical journalist, tried to organize a revolution within a revolution to secure the true equality ordained by nature. In England William Godwin, taking his stand on the immutable laws of the universe, rigorously pushed reason into condemning not only property but even the state and marriage. Such scandalous utterances, however, were only tiny specks in the sky. As the old century was dying, property seemed to be resting on a firm ground of support.

The law. For the Enlightenment, the law under which men actually lived required urgent consideration. There could hardly have been a greater contrast between the law of the courts and that which reason allegedly perceived in nature. Once Grotius had given his definition, the law of all lands was summoned to the supreme tribunal of reason. Man-made law, said the indictment, was valid only insofar as it conformed with nature's law. Even the famous jurist William Blackstone, steeped in the English common law, accepted the argument. On the Continent, where there was little ground for contentment, the spirit of legal thought was uncompromising; the law, said the zealots, was not made by edicts and legislative enactments but was discovered by right reason. The instrument of reform was a rigorous code, a summation of legal wisdom that could be administered with mechanical precision. The outcome was partial victory in the Prussian and Austrian codes and in the Code Napoléon.

Economic theory. Another aspect of human affairs that the Enlightenment sought to judge was the economic. The practical background was a long history of extensive state intervention in economic pursuits—mercantilism—and a recent history of middle class dissatisfactions. In France the wars of Louis XIV had wrought great misery, and even before the King's death, critics had dared to propose reforms. Among the first to apply rationalist criteria was the Sieur de Boisguillebert, in whose judgment state interference did violence to the law of nature, which prescribed a harmony of interests; right reasoning under the stimulus of free competition was the proper guide to economic behaviour. A half century later Louis XV's physician, François Quesnay, founded the physiocratic school on essentially the same assumptions. The physiocrats' economic man was motivated by self-interest but lived under nature's dictate of harmony; it followed, therefore, that unfettered private activity was the certain means to the commonweal.

The question of state interference

Adam Smith's *Inquiry into the nature and causes of the Wealth of Nations* (1776), a declaration of economic independence, criticized the physiocrats' emphasis on the unique importance of land but otherwise revealed Smith's kinship. Smith began with the butcher and the baker intent on their own good and went on to demonstrate the merit of "the system of natural liberty." Sharing the British dislike of extremes, he was not dogmatic; but behind his array of observations and utilitarian arguments there lurked the old rational conviction about "the great Director of Nature" leading men "by an invisible hand" to serve a wider cause than their immediate own. Smith had no illusions about the way men conducted their business affairs, and he wanted the state to protect the weak against the strong. Nonetheless, his bias was for the free individual determining his own destiny. Smith spoke at the right moment and became a revered teacher of the following generations.

The meaning of history. The Enlightenment's thought about history began with a dilemma. The basic rationalist values were timeless, and the practical purpose of examining the world's past experience was to strip away the accidents of time and place and lay bare Hume's "constant and universal principles of human nature." Another service from studying history was required: when real men were so far removed from the man of reason and nature, it was the task of the historians to explain the causes of the gap. To their labours most of the students of the past brought a point of view that suggested the answers in advance. As a result, numerous historical works glossed the theme of the evil consequences of religious and political benightedness. The histories of such as Voltaire, Guillaume Raynal, Hume, and Edward Gibbon, who recorded the triumph of barbarism and religion, were heavy guns in the arsenal of the Enlightenment. They were not, however, simply camouflaged propaganda. The Reformation had inspired solid historical scholarship, even if its purpose had been partisan, and the grave scholars of the 17th century had built on it to amass documentary materials and to broaden the skills of criticism. The luminous names of the 18th century had a surer grasp, and their masterpieces—for such many were—made Europe conscious of the pleasure and the importance of reading history.

This quietly intruding sense of historical change prompted reflection, not only on the relation of the Enlightenment to the past but also on what mankind could expect in the future. Earlier speculations were not useful guides for the latter. The ancient world had generally thought in terms of cycles in which there was a deterioration of ages from gold to iron. The Christians accepted from their predecessors the idea of cosmic cataclysm, but they would admit just one, which would come at God's own appointed time; in the meanwhile, the only significance of human history was the drama of salvation for individual souls. The Renaissance tended to return to the idea of the cycle, but it was that of youth, maturity, and decay of states after the manner of Aristotle's history of the Greek polis rather than the total upheaval of the cosmos assumed by the ancient world. The happy rebirth in 15th-century Italy was, then, merely the beginning of a new cycle. Still in the 18th century, when Montesquieu reflected on the causes of the greatness and decadence of the Romans, his interpretation continued the Renaissance approach to history.

The harbinger of a new way of thinking was Francis Bacon, the aforementioned prophet of science. The Greeks and Romans, he contended, belonged to the youth of the world, and his own generation was the heir of all accumulated knowledge. It was a mistake to be chained to the hopelessness of the notion of the cycle; by the proper method of inquiry, man could move on to great benefits through the conquest of nature. It was significant that whereas Thomas More had constructed his *Utopia* in a remote setting, Bacon put his *New Atlantis* in the future.

The idea of progress

Bacon had been thinking in terms of useful knowledge, and he found the Greeks and Romans the children and not the adults of time. Such talk was offensive to the literate community that had been raised on classical literature. For such people a comparison between the two ages had to be a comparison of their literatures, and for those

spiritually suckled on Homer and Virgil there was no comparison. By the late 17th century, brazen writers began to defend the literature of their own time, and for a whole century cultural Europe resounded with the cut and thrust of the warfare between the "ancients" and the "moderns." The battles raged on the subjective field of taste, and a clear-cut victory was manifestly out of the question. Long before the end of this contest Bernard Fontenelle, permanent secretary of the Académie des Sciences, had shifted the ground from letters to scientific achievement, and few could deny him his claim for progress up to his day. But more significantly, Fontenelle was primarily concerned with the future, and for its course he detected a law of progress that would guide man's destiny to still better things.

As social dissatisfaction, and hopes, mounted late in the 18th century, the idea of scientific and intellectual progress quietly slipped into a belief in the general progress of mankind, moral no less than material. Hardheaded realists like Voltaire were too oppressed with the continuing perversities of the world to fall into an easy optimism; knowledge could be used for bad purposes as well as for good, and future improvements were dependent on the role of sound reason. Such a qualification, however, had little restraining value for those whose vision of the future had become assured. Young Anne-Robert-Jacques Turgot, the future minister of Louis XVI, surveyed all human history and found it a record of the progress of the race toward perfection; even in ages of apparent barbarism, men and society had moved, however slowly, in the right direction and in the future would continue to do so under a law of acceleration. Turgot's friend the Marquis de Condorcet converted these calm anticipations into fervent prophecy. Hiding in 1793 from the revolution that he had joyfully served, Condorcet hastily composed his *Esquisse d'un tableau historique des progrès de l'esprit humain.* He saw mankind facing its tenth and culminating epoch—a new day in which the steady march toward perfectibility could not be stopped. The time would come, said this man under sentence of death, when the sun would shine only on free men who knew no master but their reason.

The end of the Enlightenment. A chronology of the disintegration of the Enlightenment is impossible. For all of its belief in eternals, it was from the beginning a flow of ideas in which emphases shifted and trains of thought broke under the impact of new evidence and new insights. Newton's physics and Locke's epistemology were absorbed into the uneasy synthesis, but eventually they bore a heavy responsibility for the decay. Natural law as perceived by right reason was, for many of the latter-day lovers of science, sinking into a bloodless cliché, while the pleasure-pain doctrine built up its empire of utilitarianism. At the high noon of religious rationalism, John Wesley cried defiance with his emotional Methodism. Nature, once considered a synonym of reason and visible proof of the existence of God and his benevolence, broke up into something to be studied with scientific objectivity and something to be enjoyed in romantic indulgence. Not by any means the least blow to the Enlightenment's timeless verities was the revolutionary growth of historical mindedness, of the turn toward rethinking all mortal affairs as well as scientific questions in terms of ceaseless change. In the dying days of the Enlightenment the Western mind was setting out on the road that led to Edmund Burke and Hegel and to Darwin and Marx.

Later times have ridiculed a naïve epoch, but the Enlightenment's moral and intellectual significance is high because it tried to rethink in new idioms the oldest values and beliefs in the history of Western civilization.

(Da.H.)

Revolution, reaction, and nationalism, 1789–1871

CHRONOLOGY OF THE REVOLUTION AND THE 19TH CENTURY

The French Revolution was more than a sequence of events in France between 1789 and 1795. It was one of the greatest upheavals of modern history, complex and far-reaching in importance for the whole of European civilization. It was part of a wider trend of democratic movements affecting many other countries, including North America, during the later 18th century. Modern historians see it as a continuing process, beginning in the 1770s, comprising in 1787–88 a reaction of the privileged classes of nobles, clergy, and officeholders against Louis XVI and his ministers, and passing through several phases between 1789 and 1799, with later reverberations throughout the 19th century. Events between 1789 and 1795 are the central, especially dramatic, act in a larger drama.

The Revolution in France. The importance of the meeting of the States General in May 1789 was that it brought to one focal point all the political and social conflicts of the ancien régime in France. From the only nationally representative body known to France, though it had not met since 1614, the King hoped to find an escape from his chronic insolvency. The ecclesiastical, lay, and judicial aristocracy hoped, by means of the traditional method of voting by separate estates, to block the reforms that they feared. The Third Estate, especially the middle class of officials, traders, and professional people, welcomed this opportunity to ventilate grievances and demand reforms. From this tangle of conflicting expectations derived a series of constitutional and political crises. Nobody wanted a revolution; at first nobody wanted to overthrow the monarchy and set up a republic. The hope was for peaceful changes to satisfy the various sectional interests. Yet within the next four years the surge of events and circumstances swept away first the privileges of the aristocracy and the church and then the monarchy; produced civil war and the Reign of Terror; engaged France in war with its neighbours; and spread revolutionary ideals to the rest of Europe. These unforeseen and unintended developments, the classical French Revolution, came about because of the interplay of conflicting ideological and social forces.

The first group to emerge, the enlightened liberal Constitutionalists led by such men as the Marquis de Lafayette, Bishop Talleyrand, and Abbé Sieyès, triumphed in June 1789, when the Third Estate constituted itself as the National Assembly, and in July, when it undertook to remodel the constitution. The Assembly wanted a limited constitutional monarchy on the English model. But it was soon submerged by other more violent elements in French society. The Paris mob scored a symbolic victory with the fall of the Bastille on July 14. The peasants indulged in local frenzies of destruction in the Great Fear of July and August. There were revolts in provincial towns. There began a flight of émigrés from France. When representatives of the aristocracy and clergy, on August 4, voluntarily surrendered feudal rights and privileges, they were prompted as much by fear as by generous idealism.

The forces of violence drew strength from economic conditions. The years 1787–89 were a time of bad harvests and severe food shortages, of high prices, bad trade, and unemployment. There was a slump in the textile industries. The peasants, for two decades and more, had suffered from a long-term inflationary trend that forced down their standard of living. Economic distress, deepened by the upheavals of currency, civil strife, and war in the 1790s, was the background to the political turmoil; and it, too, had its roots in the years before 1789.

The Terror. There arose within the Assembly a more extreme democratic movement, headed by the Jacobins, demanding the rule of liberty and equality and asserting the principle of "sovereignty of the people." The new constitution proclaimed in September 1791 was in essence a parliamentary constitutional monarchy; but by then the royal family, by open support of the émigrés, had destroyed the confidence between king and parliament necessary if such a regime was to work. The King was virtually held captive in Paris, and in April 1792 war was declared on Austria. The emergencies of war led to the prison massacres of September 1792 and the declaration of a republic. Louis XVI was executed in January 1793. A new assembly, known as the Convention and dominated by the Jacobins in the course of 1793, instituted the Reign of Terror under the dictatorial rule of Maximilien Robespierre. The Terror swelled the tide of émigrés, who

now included, like the victims of the Terror, people drawn from all classes. (D.Tn.)

The Revolutionary wars. Meanwhile, Revolutionary France and the monarchical powers of Europe had been engaged in a war of nerves, trading threats of increasing asperity. Direct responsibility for going beyond this war of nerves lay with the bloc of Girondin deputies in the Legislative Assembly who deliberately initiated a policy to provoke Austria into intervention and war. They were not alone in wishing that course to be followed. They themselves counted on a victorious war that would unmask what they called "the treason" of Louis XVI and Marie Antoinette and give them control over French policy. Military leaders like the Marquis de Lafayette, Louis, comte de Narbonne, and Charles-François du Périer Dumouriez, who also desired war, did so out of their conviction that France would be defeated and a grateful nation would then turn to them.

The Holy Roman emperor Leopold II, while still holding to his tactic of verbal denunciation of "the pernicious sect of Jacobins," nevertheless tightened his military alliance with Prussia, while the Prussian king Frederick William II went further and drew up a plan of campaign. The Girondins then presented an ultimatum to the Emperor, summoning him not only to renounce the Austrian-Prussian alliance but also all other treaties that threatened "the sovereignty, independence, and security of the nation." They counted on Leopold's rejection; and, when the response of Leopold's son and successor, Francis II, was held "unsatisfactory" on April 20, 1792, France declared war on "the king of Bohemia and Hungary." The deputies defined the war as "the just defense of a free people against the unjust aggression of a king."

The early fighting against the Austrians was disastrous on the northeast front, and Prussia's entry in the war greatly intensified the military danger. As the steady Prussian advance brought their troops onto French soil and as the great fortress of Verdun fell into their hands, it seemed as though only a miracle could save Paris and the entire Revolution. The "miracle" occurred on September 20 at Valmy in the passes of the Argonne. The raw French recruits held firm against the Prussian line while their artillery fire stopped the enemy advance. Though only a minor skirmish, Valmy was in effect the decisive watershed of the war. The Prussians yielded their position and began a long retreat back to the frontier. As they retreated, the French advanced in a three-pronged offensive. Savoy fell to them, the Middle Rhine, too, and most important, the Austrian Netherlands, which Dumouriez overran.

The fall of Verdun

The problem then arose in all the conquered territory of future relations between the Gallic liberators and the pro-French native sympathizers. The solution came in a momentous declaration of policy. By the decree of November 19, 1792, the new assembly of the French Republic declared "in the name of the French nation that it will bring fraternity and aid to all peoples that wish to recover their liberty . . . ." This memorable decree threw down the gauntlet to the powers. It proclaimed an ideological crusade against the old European regime of kings and privileged order of society. A complementary decree (December 15) put teeth into the declaration of policy. It stated that in all occupied territory the system of taxation, the tithe, the titles of nobility, and all special privileges and feudal dues would be abolished, and all property belonging to local rulers was to be taken over by the agents of the French Republic. These self-styled "Friends of Liberty" would institute the new administrative system, and all who resisted them would be treated as enemies.

This mystique of a crusade, coupled with the older *politique* of conquest and annexation, laid down the bases for the war of Revolutionary France against all European states, immediately against England. Threatened by the French occupation of the Austrian Netherlands and by the opening of the Scheldt to French commerce, England prepared to fight. The Convention declared war on February 1, 1793, including the United Provinces in the declaration. A month later Spain, too, was at war, then (in rapid succession) Sardinia and Naples, the lesser Italian States, Denmark, Sweden, and the Ottoman Empire. England contracted treaties of alliance and subsidy with the member states of this First Coalition. From a limited war between France and Austria, the conflict had broadened into a confrontation over two conceptions of human relations.

War with England

The defeat of the First Coalition. The fighting was a succession of defeats on all fronts. With Paris again exposed to direct attack and England declaring all France under a state of blockade, the Convention rescinded the decrees of November and December. The allies refused to negotiate: they were now discussing plans for a partition.

At this desperate juncture the Convention voted the decree of the *leveé en masse* (August 23), which was designed to transform France into an armed camp. "All Frenchmen," it read, "both sexes, all ages are called by the *patrie* to defend liberty." Implemented systematically, this declaration of total war released untapped vital energy, and the tide of battle turned to victory. By the last days of 1793 the invaders were beaten back on all fronts. Toulon was recaptured from the English, a young captain of artillery, Napoleon Bonaparte, distinguishing himself in the final attack. The Spanish withdrew to the Pyrenees and Savoy was cleared.

The élan of 1793 carried the Republican troops to victories in 1794. As the armies repulsed the invaders, they took the offensive. The line of the Rhine was regained. The Republican armies overwhelmed all of Italy. The Dutch fleet surrendered. Prussia was the first of the allies to sue for peace in the autumn of 1794. The supporters of a separate peace in Prussia reasoned that to give up territory along the left bank of the Rhine and later to receive compensation within Germany by the terms of a general European peace was a satisfactory end of the fighting.

For France the Peace of Basel (April 1795) was fully acceptable: it kept open the line of armed advance along the Rhine. To Prussia it brought the end of war and the highly advantageous status of being a neutral. By the Treaty of The Hague (May 1795), the United Provinces was compelled to cede strategic bits of territory and pay a heavy war indemnity. The reorganized Dutch government, called the Batavian Republic, was then bound to France by an alliance that drew the Dutch and their great resources into the war against England. A month later, in June, it was the turn of Spain, which had to recognize the French Republic and cede to it the Spanish part of Santo Domingo (Hispaniola). The final step taken by the French annexationists (in October) was to vote the incorporation of the Austrian Netherlands as an integral part of the territory of the Republic. The language of the French crusade to liberate enslaved peoples was kept in all these transactions; but the deeds of the liberators resembled nothing so much as the play of old-fashioned power politics.

Annexation of the Netherlands

The government of the Directory, the conservative government that came to power in France in 1795 after the Terror had run its tragic course, set its sights on a general European peace. The two allies still in the field were also eager for a comprehensive settlement, but they were not ready to accept the pre-established terms set by France. So the Directory prepared to continue the fighting. Since victory on the seas was unlikely, the French plan of campaign called for limited naval operations. The main thrust was to be on the Continent. The major attack was directed against Austria, where in a broad pincer movement two great French armies were to converge upon Vienna. A minor supporting campaign was envisaged in Italy, where the assignment given to Bonaparte, who was in command of the Army of Italy, was to conquer Lombardy. The ultimate political aim of these military operations was to compel defeated Austria to retain Lombardy in return for its acceptance of France's retention of the annexed Austrian Netherlands.

It was not only the military genius of Bonaparte but also his views concerning the territorial settlement that completely upset the plans of the Directory. As he was winning his resounding victories in northern Italy in 1796 and 1797, his thinking ranged far beyond the objectives of his government. He agreed that the conquered Rhineland would be retained. He would not, however, acquiesce in installing Austria in conquered Lombardy. He intended

to retain French military control over the territory that he was reorganizing as the Cisalpine Republic and to use it, along with French-occupied territory on the Adriatic and in the Ionian Islands, as a stepping-stone toward Egypt. For he planned to turn next to the east and strike at England by severing its imperial communications.

With such projects in mind he carried on negotiations with Austria in the summer of 1797 that eventuated in October with the Treaty of Campo Formio. Austria recognized the French annexation of the Austrian Netherlands and agreed (in secret articles) to cede approximately two-thirds of the imperial territory on the left bank of the Rhine, provided that this cession were ratified at an imperial congress to be held at Rastatt. The ousted German princes were to receive compensation elsewhere within the empire. Prussia, however (and this, too, in a secret article), was not to share in the dismemberment as the Treaty of Basel had stipulated. In Italy, Austria received Istria, Dalmatia, and the Venetian mainland in return for the loss of Lombardy. The Cisalpine Republic was formally recognized by Austria. The Ionian Islands that Bonaparte was so anxious to gain went to France.

Treaty of Campo Formio

Such were the main provisions of Campo Formio that the Directory reluctantly ratified. It was an imperialistic peace, a sorry end to the crusade for liberty. The French statesman Emmanuel-Joseph Sieyès spoke wisely when he said: "This treaty is not peace; it is a call to a new war."

The Second Coalition, 1798–1802. The Peace of Campo Formio was an interlude. While Bonaparte went off on his romantic odyssey to wrest Egypt from the Porte (Turkish government), in his absence the agents of the Directory mercilessly victimized the satellite republics set up by the military commanders. Threatened again by all these acts, Austria, England, and Russia formed the Second Coalition (1798), to which the Porte, Naples, and Sardinia adhered. Though in the early months of 1799 the allies pushed back the French on several fronts and by mid-year republican defenses were crumbling, the allied victories were transitory. By early fall, when Bonaparte returned from his Egyptian fiasco, French victories had dispelled the military danger. France was secure against invasion and could negotiate for peace from strength.

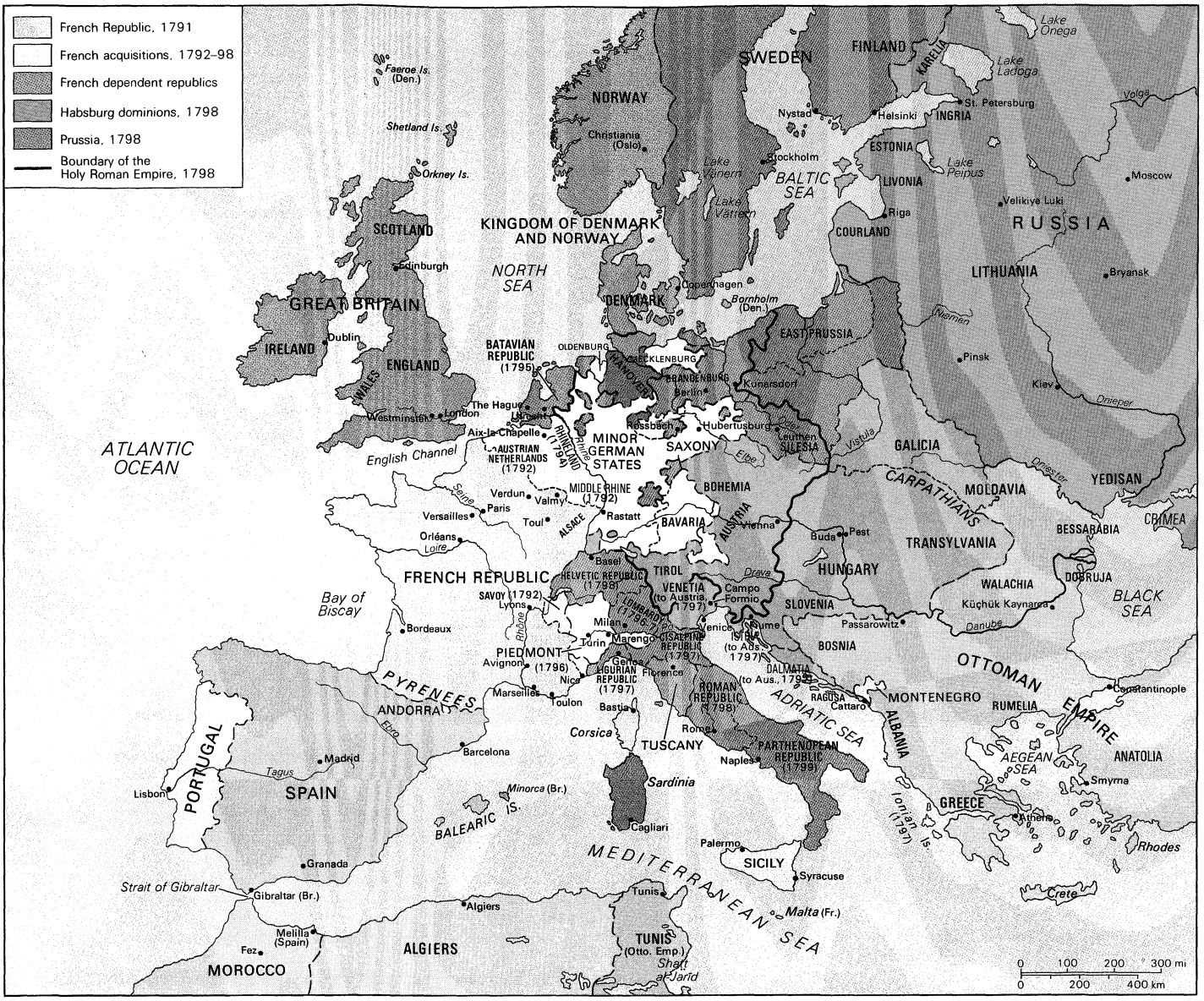
The great majority of Frenchmen yearned for peace, but not Bonaparte. He needed a major military victory to consolidate the power he had seized by the coup d'etat of Brumaire in November 1799. He drew up plans for the campaign of 1800. Once again, as in 1796 and 1797, his spectacular victories in Italy were decisive. And again, as four years earlier, Austria asked for peace. Negotiations ended with the Treaty of Lunéville (February 1801). The salient terms of Campo Formio were reaffirmed. The piecemeal ousting of Austria from the Italian peninsula continued, but more important than its expulsion from the Duchy of Tuscany was the Habsburg recognition of the new Italian republics as sovereign powers.

The Treaty of Lunéville

Only England remained at war, and the English government had to yield to the insistent clamour for an end to the struggle. Peace talks were begun a month after Lunéville. The preliminaries of October 1801 were converted into the definitive Treaty of Amiens on March 27, 1802. The provisions were harsh. Egypt was restored to the Ottoman Empire. Both Malta, which went to the Knights of St. John, and the Cape of Good Hope were restored under conditions that made their future security precarious. Save for Ceylon and Trinidad, England restored its maritime conquests. Moreover, England tacitly accepted the main territorial changes on the Continent as they had been agreed upon at Lunéville. Amiens contained no provisions

Maps © George Philip & Son, Ltd. 1970; all other material © Cambridge University Press 1970 from *The New Cambridge Modern History*, volume XIV, "Atlas," edited by H.C. Darby and Harold Fullard

Sites associated with the French Revolutionary and Napoleonic wars.



The expansion of the French Republic, 1791–98.
From *Grosser Historischer Weltatlas*, vol. 3, *Neuzeit* (1972); Bayerischer Schulbuch-Verlag, Munich

for a favourable commercial treaty that would reopen the continental markets for English goods.

In the circumstances, a chorus of recrimination fell upon the government. England obtained, so it seemed, only what Bonaparte gave it: peace on his terms. After ten years of war England's great rival and competitor emerged as the most powerful state in Europe. "One may state without the slightest exaggeration," Talleyrand was to write later, "that at the time of the Peace of Amiens, France enjoyed abroad, thanks to its military supremacy, such power, glory, and influence that the most ambitious person could desire nothing more for his country." Bonaparte, however, desired more. (L.Ge.)

Napoleon in power: Lunéville to Tilsit. Neither the Treaty of Lunéville nor that of Amiens contained the seeds of a long-term peace—the first, because Napoleon gradually broke all its terms in the interests of French aggrandizement; the second, because both parties had signed it out of temporary exhaustion and did not regard it as satisfactory.

Consolidation and strengthening of French power

After Lunéville Napoleon, who became first consul for life in 1802 and emperor in 1804, lost no time in consolidating and strengthening French power, particularly in Germany and Italy. In February 1803 the Diet of Regensburg approved French plans for the reorganization of Germany, which led to the formal death of the Holy Roman Empire in 1806. This reorganization, to some extent made necessary by French expansion to the Rhine, led to the suppression of ecclesiastical states and to considerable gains by the more important lay princes. In the process, French patronage became the controlling influence in Germany. The United Provinces—now the Batavian Republic—was bound yet closer to France by late 1801. In 1802 the Cisalpine Republic became the Italian Republic, with Napoleon as its president. In 1802 Piedmont was also incorporated into France. In 1803 Napoleon mediated in a civil war in the Helvetic Republic (which then became Helvetia) and followed that action with a new constitution strengthening French influence there. Later in 1803 the French entered Hanover and took it under French protection. (Small wonder that the governments of other European countries felt that there was no predictable limit to the ambitions of the first consul.)

After Amiens Britain, although nominally at peace with France, found no basis of mutual trust for the future. Apart from Napoleon's continued aggression on the Continent, he was already imposing restrictions on British trade and, in 1803, was suspected of encouraging rebellion in Ireland. Napoleon complained that the British had not surrendered Alexandria or Malta (he was correct) and that Britain was sheltering and encouraging French émigrés. It was not therefore surprising that

the two powers found themselves at war again in May 1803.

The war against Britain, to Trafalgar. At this stage Britain had no allies and would have none for nearly two years. Hostilities with France on this one-to-one basis revealed more clearly than before the stalemate that could be reached in a situation in which a great land power was opposed to a great sea power with neither able seriously to oppose the other in its own element. Napoleon could assemble 100,000 men at Boulogne—a far larger and more experienced army than anything Britain could oppose him with—but had no way of getting his soldiers into action. The Royal Navy could prevent the French Army from crossing the English Channel but had no means of attempting, let along ensuring, the defeat of that force.

The British, first with Henry Addington and then (from May 1804) with William Pitt as prime minister, set about sweeping up the remaining French colonial possessions. St. Lucia, Tobago, and Guiana were taken in 1803 in the Western Hemisphere; in the same year French influence in India was further weakened by the defeat of its friends there. Then, beginning in mid-1804, Pitt began his attempts to divert and thus weaken Napoleon by rebuilding a continental coalition. In June 1804 he had reached an understanding with Russia that was converted into a formal alliance in April 1805. That August Austria joined again, and the Third Coalition was in being.

The Third Coalition

Meanwhile the war at sea continued. Napoleon attempted to stop Britain's trade with the Continent: he seized Cuxhaven for this purpose, only to meet with a British blockade of the mouths of the Elbe and Weser, which considerably damaged Napoleon's friend Prussia. The French also continued their attacks on British seaborne trade by means of privateers. The most important aspects of this war between France and Britain, however, were French preparations for an invasion of Britain by a cross-channel attack and Britain's blockade of the ports of France and her allies.

The naval defeats of the 1790s had not persuaded Napoleon against building new fleets. Nevertheless, his resources were decidedly inferior to those of the Royal Navy in all classes of vessels, and his own admirals preferred a defensive "fleet-in-being" strategy combined with occasional brief sorties. For Napoleon himself the chief hope lay in a possible invasion; during 1803–04 troops were assembled in the Boulogne area, together with 2,000 transports. His plans for invasion varied between one employing a heavily protected force and one conducted simply in small boats, on their own, while French squadrons provided diversions elsewhere and drew away the blockading British squadrons. The right combination of successful diversion, weather, and tide was never found. Napoleon had broken camp at Boulogne and begun a long march to the Danube against Austria seven weeks before Nelson won a great victory at Trafalgar.

For Britain the basic response to invasion was the blockade, both close and distant. British squadrons were spaced off major naval ports—from the Texel along the French and Spanish coasts off Brest, Ferrol, Cádiz, and Toulon—to prevent French and Spanish squadrons from putting to sea with supporting frigates for intercommunication. Because invasion of England or Ireland was the chief danger, it was established that squadrons would fall back to a position off Ushant if one or more powerful squadrons broke the blockade and thus threatened to act as an invasion escort. (Allowing for all the differences between warfare at sea and on land, this was not dissimilar to some important aspects of Napoleon's own strategy.) A constantly successful blockade, however, was too much to expect, and both Napoleon and the Royal Navy knew that. British forces had to be split into small squadrons. Long periods at sea meant ships would sometimes be absent for refitting without adequate replacements, and the uncertainties of the weather sometimes helped the blockaded. Once the French were at sea, the Royal Navy, despite all its precautions, was almost inevitably caught in some degree of uncertainty as to where the French would go.

On March 31, 1805, Adm. Pierre-Charles Villeneuve, commanding the French squadron in Toulon, evaded Nelson's blockade and broke out of the Mediterranean, heading for the West Indies. Meanwhile, the other major French squadron was blockaded in Brest. Nelson chased Villeneuve to the West Indies and back, while other Royal Navy squadrons swarmed in the Channel to prevent a French concentration that could be used to convey an invasion force. Villeneuve returned and took refuge in Cádiz; and Nelson, convinced that the French had plans for the Mediterranean, followed him there. On October 20 Villeneuve, who could not stay in Cádiz because he lacked supplies and facilities, left there and was caught by Nelson off Cape Trafalgar. The French had 33 ships of the line to 27 of the British; but the latter had a tactical plan, and Villeneuve lacked one. In the Battle of Trafalgar (October 21), 18 French ships were totally lost and the rest fought their last action. This was the last battlefleet action of these wars, and the danger of an invasion of England was ended.

End of the war at sea

The campaign against Austria, 1805. As a result of its defeat in 1800–01, Austria's influence in Germany had been greatly weakened. The old Holy Roman Empire in practice no longer existed, a fact recognized by Francis II himself when he was proclaimed emperor of Austria in August 1804. During this time Britain and Russia were drawing closer together; Alexander I of Russia broke off diplomatic relations with France in the autumn of 1804 and signed a treaty of alliance with Britain in 1805 designed to drive the French from northern Germany, Holland, Switzerland, and Naples. Austria, still weakened by the effects of past defeats, hung back. But, when Napoleon had himself crowned king of Italy in 1805 and then incorporated Genoa and the Ligurian Republic into France, Austria finally entered the Third Coalition with Britain and Russia in August of that year. Napoleon gave up his plans for invading England to deal with Austria and Russia. Writing to the French statesman Talleyrand on August 13, Napoleon announced,

> My course is settled. I shall raise my camps [*i.e.*, at Boulogne] . . . by 23rd September I shall be in Germany with 200,000 men, and have 23,000 men in Italy . . . I shall march on Vienna and compel Austria to sue for peace.

The campaign conformed very nearly to that forecast.

The army corps system

Napoleon was faced with three problems. First he had to keep the Austrian Army separated from the Russians and persuade it to fight on its own; he planned to advance from the Middle Rhine on Vienna, if possible cutting the Austrian line of communication with Vienna somewhere east of Ulm, thus catching the Austrian Army with an attack from its rear as he had done at Marengo. The second problem was to move 200,000 men across Europe from the Channel coast to the Middle Danube in the space of about six weeks and yet retain the ability to concentrate this vast army, if necessary within a day or two, at the critical time and place; he planned to make maximum use of the army corps system, the corps having become a unit of all arms, capable both of combining with other corps in a major battle and of operating on its own, even if only in a holding or defensive capacity, against a major enemy force. The final problem was that of provisioning so many men during so rapid an advance.

Regarding provisions, each corps had its own wagon train, and magazines were sometimes established at the most important towns en route. That provisioning system, however, did not always provide enough; if magazines were not available, French armies, using this blitzkrieg strategy, had to live off the land by requisitions. Each corps was allotted a requisition area strictly in relation to its prescribed line of advance. Commenting on this later, Napoleon wrote,

> We have marched without magazines; circumstances have forced us to do this; but although we have been continually victorious and found vegetables in the fields, we have nevertheless suffered a great deal, and, in a season when potatoes cannot be had from the fields a lack of magazines would lead us into the greatest misfortune.

Gen. Karl Mack, commanding the Austrian Army of about 85,000 men, played into Napoleon's hands. At the beginning of September he advanced westward through Bavaria toward Ulm without waiting for his Russian allies.

At Ulm he hoped to command the line of the Danube, with his front and right protected by the Black Forest. By this advance, however, he had greatly lengthened his line of communications; and, even as late as October 9, he was still separated from his Russian allies, who were at Linz, by nearly 200 miles (322 kilometres).

Austrian defeat

Meanwhile, Napoleon's advance guard crossed the Rhine on September 24, and his army advanced on the Danube from the north and west on its prescribed corps fronts. On October 6 his forces were in the area of Donauwörth-Neuberg and then crossed the Danube, leaving Gen. Michel Ney with 40,000 men to cover Austrian escape routes to the north. Napoleon then continued to envelop Mack in the south. Mack made two unsuccessful attempts to break out of the ring and then capitulated at Ulm on October 20 with 50,000 men. (If great victories consist of beating an enemy at a minimum cost in actual fighting, then Ulm was one of the greatest of Napoleon's victories.)

Austria was not yet completely defeated. On November 13, three weeks after Trafalgar, Napoleon entered Vienna; but he still faced a combination of Russians commanded by Gen. Mikhail Kutuzov and two Austrian armies from Italy and the Tirol. Moreover, Napoleon himself was now operating at the end of a long line of communications. Had the Allies waited—they were, after all, on their own ground—time would almost certainly have been on their side; but some of them, particularly Tsar Alexander, were impatient and chose to attack. On December 2 Napoleon routed them at Austerlitz, a battlefield victory even more complete than the victory at Ulm.

In the resulting Peace of Pressburg, signed at the end of 1805, Austria ceded its last Italian possession, Venetía (Venice), to France, gave the Tirol to Bavaria, and recognized the rulers of Bavaria and Württemberg as independent kings. Already cut off from the Rhineland, Austria was excluded from Italy and gravely weakened in southern Germany.

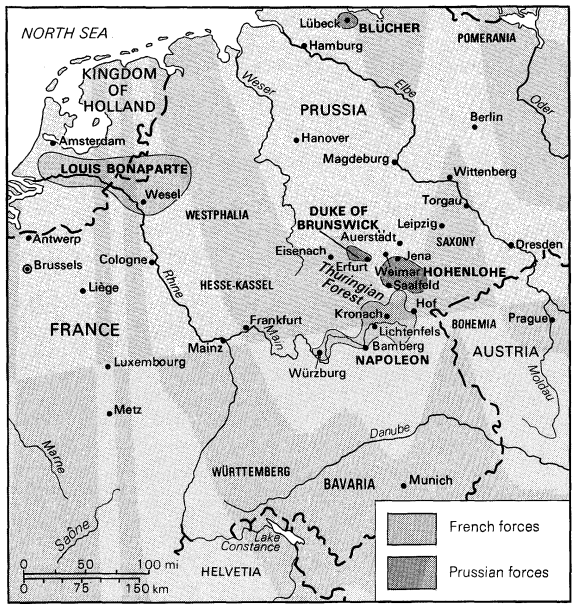
The campaigns against Prussia and Russia, 1806–07. During the 1805 campaign Prussia had been tempted to join the Allies; such a move might have threatened Napoleon's left flank as he moved along the Danube Valley and perhaps even have denied him victory. Frederick William III, however, delayed acting until after Austria had been defeated, thus weakening his position when he himself was attacked.

Throughout 1806, relations between France and Prussia steadily worsened. Despite his continued gestures of friendship toward Prussia, Napoleon appeared determined to weaken Prussian influence in northern Germany, and his independent negotiations with Britain over Hanover underlined this. Furthermore, after Austerlitz Napoleon kept the bulk of his army in Germany, stretched out along the line of the Main River; by October 1806 he had about 200,000 men on a line running from Lichtenfels through Bamberg to Würzburg. The Prussians regarded this troop concentration as a threatening gesture; Russia, too, was suspicious of Napoleon's intentions. On September 26, Frederick William III of Prussia sent an ultimatum to Napoleon demanding that the French withdraw west of the Rhine and agree to the formation of a north German confederacy under Prussian leadership. Napoleon rejected the ultimatum. On October 8 Prussia, in alliance with Russia, declared war on France.

The overall stategic situation was not unlike that in 1805. With promises of help from Russia and from Sweden and Britain too, there was much to be said for the Prussians staying on the defensive on the line of the Elbe and waiting until help arrived. The Elbe formed the great natural defense of Prussia against attack from the west, and the main passages across the river were guarded by the fortresses of Magdeburg, Wittenberg, Torgau, and Dresden, which also controlled the principal routes into Berlin and East Prussia. The one advantage of advancing west of the Elbe was that Saxony and Hesse-Kassel might join with Prussia if protected in this way but probably not otherwise. Indeed, Saxony could help Napoleon directly by offering him an unopposed passage of the Elbe at Dresden.

Prussian advance

In fact the Duke of Brunswick, in command of the Prussian army, advanced west across the Elbe, then divided his forces in two—the main part under his command concentrated in the area of Erfurt, the other part south of him, commanded by F.L. Hohenlohe-Ingelfingen and concentrated between Hof and Saalfeld in the Upper Saale Valley. These two forces were separated by the difficult terrain of the Thuringian Forest.



The location of troops on October 8, 1806.

Adapted from V. Esposito and J. Elting, *A Military History and Atlas of the Napoleonic Wars* (1964); Praeger Publishers

For Napoleon the possible routes from the Rhine into Prussia were either from Wesel across Westphalia and Hanover, from Frankfurt via Eisenach and Weimar to Leipzig, or from Frankfurt via Bamberg to the south of the Thuringian Forest and thence either by Leipzig or Dresden. He treated the first route, the most northerly, as cover against possible enemy retaliation, leaving a corps under E.A. Mortier to link up with a force under Louis Bonaparte, whom Napoleon had made king of Holland. This tactic meant that, with a northern escape route safe, he could risk his other lines of communication. He then planned his advance along the third route, telling Louis weeks in advance that he intended to be in position to attack the Prussians on October 12. He left Mainz on October 1, used magazines at Würzburg and Kronach on the way, and crossed the frontier of Saxony on October 8. His rapid deployment took the Prussians by surprise, and Hohenlohe promptly began to retreat northeastward toward the Elbe when contact was established on October 12, as Napoleon had predicted. On October 14 Napoleon overwhelmed Hohenlohe at Jena, while Louis-Nicolas Davout, sent round the flank to block a Prussian retreat, encountered the main Prussian force at Auerstädt and defeated them also. The battle had not gone quite in detail as Napoleon planned, but in general execution he had dictated the critical time and place and had made the Prussians conform to his own movements.

Prussian retreat followed. On October 25 Napoleon entered Berlin, and ten days later the Prussian rear guard under Gebhard Leberecht von Blücher capitulated at Lübeck. Prussian resistance was at an end. Russia, however, was still in the war; in November Napoleon invaded East Prussia and then Russian-controlled Poland, occupying Warsaw a week before Christmas. Marching still farther eastward, he met the Russian Army under Leonty Leontyevich Bennigsen at Eylau (February 8, 1807), fighting a bloody and indecisive battle. In mid-June, despite his long line of communications and his earlier losses, Napoleon routed Bennigsen at Friedland, driving the Russians back across the Niemen while he himself entered Königsberg.

The Treaty of Tilsit

The war was over, and peace was signed at the Treaty of Tilsit on July 9. This treaty really dealt with Prussia rather than Russia. Prussia lost all its territory west of the Elbe and the part of Poland it had gained most re-

cently. Its territory was thus halved and its army reduced to 40,000 men, one-fifth of its previous size. (Prussia lost more, at one blow, than any other of Napoleon's principal enemies.)

Russia emerged virtually unscathed from the 1806–07 campaign. Indeed, it became an ally of France, and the events of the next few years were, to some considerable extent, based upon this Franco-Russian alliance.

The Napoleonic coercion of Europe, 1807–11. With the signing of the Treaty of Tilsit, Napoleon found himself at the peak of his power. So long as he and Alexander I could agree and, in agreeing, coordinate their policies, Napoleon could anticipate a time when all continental Europe would lie under his control. The Continental System that he organized in detail was the most obvious expression of his power. The first signs of decline, however, coincided with the period of ostensibly greatest power. As the expansion of French power became increasingly identified with blatant imperialism, the seeds of nationalist revolt began to bear fruit—first in Spain and then in Austria—producing movements that, in the end, broke Napoleon's power. Moreover, the Continental System was too ambitious politically, administratively, and economically, and it weakened its creator as much as it strengthened him.

The peak of Napoleon's power

The Continental System. Trade war against Britain was not initiated by Napoleon. After war broke out between Britain and France in 1793, the republic did its best to exclude British goods from the areas under its control. The Directory stepped up measures against British trade and ordered the seizure of all vessels that put in at British ports and then sailed on to France. Moreover, the theoretical inspiration of this kind of warfare was similar both before and after Napoleon became first consul. A well-established mercantilist strain in French thinking on these matters held that the best way to defeat a major trading power such as Britain was to stifle or cut off its exports and thus reduce its stocks of gold; this view was itself based on the theory that gold, produced by a favourable balance of trade, was a vital indicator of national power. Napoleon based his Continental System on these theories. The difference between him and his predecessors was a practical one—his conquests and particularly his conquest of Prussia and subsequent agreement with Russia enabled him to apply such a policy on a genuinely continental scale. Although he was a radical in his approach to methods, he remained a conservative (some would say a reactionary) in his economic theories. Finally, after Trafalgar and even more so after Britain's seizure of the Danish fleet in 1807, Napoleon had no reasonable hope of resuming normal warfare at sea; and even the threat of his privateers was severely reduced in face of an increasingly effective convoy system carried out by the Royal Navy. To that extent and whatever Napoleon's original hopes of its success, the Continental System can be regarded as a method adopted for lack of something better.

The defeat of Prussia at Jena meant that Napoleon could control a long stretch of the Baltic coast. On November 21, 1806, Napoleon issued his Berlin Decree, which placed the British Isles in a state of blockade. This action meant that trade with Britain was prohibited and that all vessels coming to ports under French control directly from Britain or its colonies would be seized. With Russia adhering to the System and with Portugal and Spain coming under French control in 1808, Napoleon could well claim that he hoped to conquer the sea by the land. The government in London retaliated with a series of Orders in Council in 1807, the object of which was essentially to bring seaborne trade with Europe as far as possible under British control by allowing such trade in neutral ships to take place only by license. This neutral trade was considerable—in 1807 about 44 percent of all trade passing through British ports was in neutral ships, and these ships were to come under British direction as well as provide a useful source of income for the treasury through the issue of licenses.

The Berlin Decree

Napoleon's reply to the Orders in Council was to take his own system a stage further. By the Fontainebleau and Milan decrees of 1807, he ordered that all neutral ships conforming with the rules laid down by the Orders in Council should be treated as enemy ships and, where possible, seized. Thus, neutrals had either to risk detention by the British or confiscation by the French; and that dilemma applied particularly to U.S. vessels. A further Fontainebleau decree of October 1810 strengthened the law against contraband and also ordered the public burning of captured British manufactures. Other decrees that year were designed to tighten up Napoleon's own system of import and export licenses.

The restrictions imposed by the Continental System and by Britain's countermeasures had serious effects upon Britain and also upon France and its continental allies and satellites. During 1808 there was a considerable drop in British exports—a fall of about 10 percent from 1806. Cotton and grain imports were also affected, leading to shortages, high prices, and unemployment. The shortage of timber hit hard at both naval and merchant shipbuilding. There was an improvement from Britain's point of view in 1809, when Portugal, Spain, and Spain's colonies escaped from French control and when the Ottoman Empire signed an agreement with Britain.

But 1811 was the most critical year of all for Britain. Its Baltic trade suffered a severe blow when Sweden was included in Napoleon's System. In that year exports to northern Europe were only one-fifth of the previous year's total, and exports to the United States were down by a quarter. Overall, British exports dropped by about a third. The trade crisis was worsened by a monetary crisis in which the pound depreciated and reserves dropped dangerously low. Many Britons talked of peace at any price; had Napoleon—who had recently agreed to the export of French grain to Britain partly to placate French farmers and partly in pursuit of his mercantilist theories—chosen to deny wheat exports to Britain at this point, hunger riots might have tipped the scale in forcing the British to make peace.

While it did produce these undoubtedly serious effects, the Continental System failed in its prime objective of defeating Britain by crippling its trade and thus ruining its credit. Napoleon's unsuccessful attack on Russia in 1812 meant that the sizable cracks in the System that had already appeared in Europe became great holes. This situation helped offset some of the effects of the outbreak of war in that year between Britain and the United States, itself a result partly of disputes about the freedom of the seas. Behind these important ups and downs of fortune, however, lay one constant factor—Britain's command of the sea. As the U.S. naval historian Alfred Thayer Mahan has pointed out, Napoleon was bound to lose a trade war against a country that, because of its vast navy and mercantile marine, could offset losses in one area of the world by opening up opportunities elsewhere. And that is precisely what Britain did. Its national debt grew enormously; but basic prosperity was founded upon increasing industrial production and trade, and Britain emerged from the war much wealthier and stronger than it had entered it.

Failure of the Continental System

The effects of the Continental System upon France itself were more generally damaging. Napoleon faced two problems. The first was how to make his allies and satellites continue to accept restrictions that were often harmful to them; in the end resentment about the System helped to turn allies into enemies. The second problem was in France, where his situation was equally difficult. It is true that he was always willing, perhaps wisely in a political sense, to sacrifice the interests of his European allies to those of France itself. This favouritism came out quite clearly in licenses granted much more freely to French traders. This preference, however, made little difference to French trade in the long run. At first some French traders welcomed the advantage conferred by the removal of British competition; but France, like other European countries, needed markets and colonial raw materials such as cotton and sugar. Customs receipts and exports fell drastically. Gradually Napoleon lost the support of the middle class, who had gained so much from the Revolution and on whose efficiency and loyalty Napoleon himself depended. Ersatz industries were an adventurous experiment, and some came to stay. By and large, however, the Continental System undermined—or certainly was thought to

undermine—prosperity, and Napoleon could not afford that.

Austrian opportunity for revenge

Austria's war of liberation, 1809. Angered by the terms imposed upon Austria after Ulm and Austerlitz and offended even more by Napoleon's further aggrandizement after the defeat of Prussia, the emperor Francis of Austria and his advisers thought they saw an opportunity for revenge in 1809. The French Army was heavily committed in Spain; and there was rising discontent in Germany, as elsewhere, from the effects of the Continental System. Moreover, the Austrian government had carried out reforms in the army after 1806 and had a military-reserve system by 1808. The tide of public opinion was moving against Napoleon; and, although there was no strong German nationalism evident in Austria or elsewhere at that stage, there was much local patriotism—for example, in the Tirol.

Napoleon was aware of events in Austria and had carefully made his plans by assuring himself of Russia's neutrality. Prussia could safely be ignored for the time being. Moreover, although he spent the winter of 1808–09 in Spain, Napoleon was back in Paris before Austria declared war on France in April 1809. By calling up 150,000 conscripts, he had 300,000 men at his command for operations in central Europe, in addition to the troops committed in Spain.

The Austrian plan was to attack the French in Italy and Dalmatia; to advance into Poland, forestalling help from Russia and the Grand Duchy of Warsaw; and to advance westward along the Danube Valley to drive back the French concentrated at Regensburg and Augsburg before Napoleon could cross the Rhine and come to their aid. The archduke Charles, in command of this last Austrian army, was much too dilatory and allowed Napoleon to join up with Davout's forward army corps. By mid-April Napoleon had concentrated his forces and advanced toward Vienna, defeating the Archduke at Eckmühl on April 22. The Archduke was forced north of the Danube, and Napoleon entered Vienna on May 13.

The Austrians remained strong both north of the Danube and farther south in the Tirol. Napoleon concentrated his attention on the archduke Charles and his army north of the Danube. On May 20 Napoleon crossed the river by the island of Lobau, and a battle took place at Essling-Aspern on May 21–22 in which, though not completely defeated, Napoleon was mauled seriously enough to retrace his steps across the river. He then fortified his position and, in early July, was reinforced by the army of Eugène de Beauharnais, which had come up from the south. On July 5 Napoleon crossed the Danube again with an army of more than 150,000 men and the next day defeated the archduke Charles at Wagram. An armistice was signed a few days later.

By the Peace of Schönbrunn (October 1809) Austria was yet further humiliated. It lost its Illyrian provinces and ceded much of Upper Austria to Bavaria and most of western Galicia to the Grand Duchy of Warsaw. Its efforts, however, had not been in vain; French troops were withdrawn from Spain. Moreover, Napoleon's greed in taking more Austrian territory for his satellites offended his friends as well as his enemies.

The Spanish rising and the Peninsular Campaign. A more prolonged and in many ways very different campaign was meanwhile going on in Spain and Portugal. Spain had been entirely subservient to French policies since 1796; and, although there was a flicker of resistance during the campaign against Prussia, the French victory at Jena brought Spain once more to heel. By then Napoleon was intent on compelling all Europe to accept the Continental System, and Portugal, for so long the ally of England, was an obvious gap in the System. By the Treaty of Fontainebleau of 1807, the Spanish government agreed that French troops could cross Spain on their way to conquer Portugal, and in November the French entered Lisbon. These policies led to deep differences in Spain. In May 1808 the king, Charles IV, abdicated, and in June Napoleon appointed his own brother Joseph as king of Spain. There had already been a popular rising against the French at Aranjuez in March 1808, and the events of that summer spread the flames of insurrection against what was regarded as French tyranny. A national rising against France was the result.

War in Spain

The importance of war in Spain was twofold. Politically it was the clearest indication to date of nationalist resistance to Napoleonic domination, even though it was not a national movement inspired by liberalism. Militarily, it afforded Britain an opportunity to undertake land operations on the mainland of Europe that, at last, were significant in relation to the operations of its allies. For the first time sea power was directly related to land power. The Iberian Peninsula was easy for Britain to get to and to use, and it was difficult for Napoleon to defend. The logistic and military basis of Britain's whole operation was Portugal, which offered a first-rate entry, via the Tagus and Lisbon, and a sound defensive position.

Napoleon himself took command in Spain late in 1808 and very nearly caught a British force before it managed to escape via Coruña. Early in 1809, however, he returned to Paris. In April 1809 Sir Arthur Wellesley (later the duke of Wellington) landed at Lisbon and took command of all British and Portuguese forces there; he had come carefully prepared and had already formed some views about the best ways to fight the hitherto invincible armies of Napoleon.

At a tactical level, Wellesley believed in the value of highly disciplined firepower delivered by an infantry line protected, wherever possible, by natural features or by such man-made features as walls; and he also believed in the importance of light infantry. Secondly, Wellesley was determined to establish an efficient supply system within his army so that it became independent, as far as possible, of army requisitions and could thus move or stay in its positions as the military situation demanded. He abhorred the French system of living off the land, and he considered that it forced the French to move when often they would have served their military purposes better by staying in position. Thirdly, he developed a highly efficient staff system, all the more necessary in a campaign with frequently divided forces; and his staff gave him, among other things, an invaluable series of maps that enabled him to move in difficult country with certainty and speed.

Wellesley's strategy

Wellesley's strategy was to use Portugal as his base, with its land frontier defended by a number of great fortresses and with a final defensive area behind the lines of Torres Vedras around Lisbon. The sea was his chief line of supply. In 1809 Wellesley (Viscount Wellington from September 1809) had some early success, driving the French from Portugal and in July defeating them at Talavera; but by the autumn he was back behind the Portuguese frontier.

In 1810, with Austria defeated at Wagram, Napoleon sent reinforcements to Spain; and the positions were reversed. By that autumn the French had captured the Portuguese frontier fortresses and then occupied all of Portugal except the area behind the lines of Torres Vedras.

From then on, Wellington was more certain of his own strength and of enemy weaknesses. In 1811 he began the process of recovering Portugal and, principally, its frontier fortresses. In May he defeated André Masséna at Fuentes d'Onoro and captured Almeida—the first in a series of defeats suffered by the French. In 1812 he captured Ciudad Rodrigo and Badajoz before the end of April, and, in July, he routed the French at Salamanca, going on to enter Madrid in August, thus compelling French forces to withdraw from southern Spain. In the autumn of 1812, however, Wellington ran into strong French resistance as he tried to capture Burgos, and once more he withdrew to winter quarters in Portugal. In 1813 he advanced again, defeated the French at Vitoria, forced them to evacuate central Spain, and began an invasion of southern France across the Pyrenees; by the end of the year he was threatening Bayonne, and in February 1814 he captured Bordeaux. The Peninsular Campaign was now complete. Wellington's advance had been fused into the general Allied attack on France that was to bring about Napoleon's first abdication.

The defeat of Napoleon, 1812–15. Although the Treaty of Tilsit had been, ostensibly, a Franco-Russian alliance and although Alexander I had refrained from embarrass-

Causes of Franco-Russian disagreement

ing Napoleon in the difficult situation of 1809, there were increasing causes of tension between Napoleon and the Tsar. Alexander was anxious to extend Russian influence southward through the Balkans, and Napoleon wanted to prevent that influence from reaching the Mediterranean. Napoleon had created the Grand Duchy of Warsaw, and Alexander was suspicious of French influence so close at hand, particularly if Napoleon had it in mind to re-establish the old Kingdom of Poland. Finally, the working of the Continental System produced as much discontent in Russia and in areas of Russian influence as it did elsewhere.

The Russian Campaign, 1812. These general causes of disagreement partly came to a head in 1810–11. In February 1811 Napoleon seized the Hanse towns and the lands of the Tsar's brother-in-law—the Duke of Oldenburg. That May, when a Frenchman, Jean Bernadotte, became crown prince of Sweden, it seemed to Alexander a French step toward encircling him—although, in fact, Bernadotte proved no friend to Napoleon. In January 1812 Napoleon issued a list of grievances against Russia to his German allies, and in April Alexander presented an ultimatum to Napoleon demanding French evacuation of Prussia, compensation for Oldenburg, and virtually the creation of a neutral zone between the two power blocs. In May, Russia concluded a treaty with the Ottoman Empire and thus freed itself from trouble in the Balkans. In July 1812 a treaty of alliance was signed among Sweden, Russia, and Great Britain.

Meanwhile, Napoleon had already decided that he would invade Russia in June, and he concentrated a vast army of about 600,000 men in the Grand Duchy of Warsaw, drawing heavily on troops from allies and satellites. About 450,000 men actually crossed the Niemen River at the beginning of the campaign and were opposed by fewer than half that number in the Russian forces.

Napoleon had the choice of three good roads leading from west to east and toward Moscow—from Kovno via Vilna, Vitebsk, and Smolensk; from Grodno via Minsk and so to Vitebsk; and from Brest-Litovsk via Kiev to Smolensk. He decided to take the first, the most northerly route, using the Grand Duchy of Warsaw as his strategic base and with his lines of communication lying back through what he thought were friendly areas of Poland and Prussia. His strategic approach, after assembling his armies behind the line of the Vistula from Warsaw to the coast, was to use his right flank as a defense and to attack on his left flank from Kovno eastward with his largest army under his own command.

The Russian forces were deployed in two armies. The larger of the two, commanded by Mikhail Barclay de Tolly, was along the line of the Niemen north of the Pinsk Marshes (*i.e.,* on the right, or north, flank) and comprised about 135,000 men. The Russian left-flank army, commanded by Pyotr Ivanovich Bagration, was to the south of the Pinsk Marshes in Volynia and had a smaller force of about 50,000 men. From the start, the Russian commanders were well aware of the risk of separation and envelopment and were determined on a strategy of withdrawal upon converging lines. Writing to the King of Prussia months before the campaign began, the Tsar said,

Russian strategy

> The system which has made Wellington victorious in Spain, and exhausted the French armies, is what I intend to follow—avoid pitched battles and organise long lines of communication for retreat, leading to entrenched camps.

Quite apart from avoiding the risk of defeat on the battlefield, such a strategy was bound to impose a heavy strain on Napoleon's forces in their search for food and fodder and in a constantly lengthening communications system in an enemy country. (When Clausewitz wrote some years later that one of the main strategic advantages of the defensive was the ability to make use of the goodwill of the local population, he almost certainly had this particular campaign in mind.)

Two further points of general strategic importance can be made. First, the roads in areas in which Napoleon planned to operate were mostly of poor quality and distinctly inferior in standard to the road systems of western and central Europe. Second, it was normally possible to continue campaigning in western Europe until late in the year without serious inconvenience, but, in Russian winter conditions, victory consisted of survival rather than of defeating the enemy forces; and in the search for survival Napoleon's troops were at a great disadvantage.

On June 24–25 the entire French left wing crossed the Niemen at Kovno almost unopposed. Napoleon expected soon to be behind Barclay, but, when he arrived at Vilna on June 28, he found that Barclay had moved eastward and escaped the net. Moreover and despite the summer weather, logistic problems were already retarding the speed of Napoleon's movement. Although continuing his advance against Barclay, Napoleon turned his attention southward to Bagration, wrongly supposing that the latter was retreating in a northeasterly direction and planning to prevent the two Russian armies from uniting in the area east of Vilna. But Bagration had gone southeastward instead via Minsk, thus escaping in his turn. Moving again to the north, Napoleon tried to encircle Barclay both at Vitebsk in late July and at Smolensk in mid-August. On both occasions the Russians slipped out of the net again and retreated toward Moscow.

By then the French advance was running into serious supply difficulties, which robbed Napoleon of mobility and imposed on him a blunt frontal attack strategy in place of envelopment. After a bloody action at Borodino on September 7, the Russians continued their retreat; and the pursuing French at last reached Moscow in mid-September. There they found the city devastated by fire but showing no sign of capitulation. After a month in Moscow, Napoleon himself began to retreat, and, although he was at no time defeated by the Russians, he was forced, by the fact that they followed fast on his heels and preyed on his flanks, to follow the route by which he had earlier advanced, but now substantially denuded of supplies. This fact and the severe weather conditions took their toll. Napoleon reached Smolensk on November 8 and crossed the Beresina on the 17th. On December 5, hearing rumours of a conspiracy against him in Paris, Napoleon left his army and made for home. That army was now experiencing the Russian winter; by the time it reached and crossed the Niemen in mid-December, the temperature was −30° F (−35° C), and only about 30,000 men had survived out of the 600,000 Napoleon had so hopefully assembled the previous summer. At the Niemen the Russian pursuit stopped, and the remaining French forces continued their wretched withdrawal unopposed.

The French in Moscow

Napoleon's military power was not annihilated, as the campaigns of 1813–14 were to demonstrate. But he had lost invaluable troops; his reputation was at last open to question; and his enemies had been given a morale boost that it would take many French victories to destroy. The Russian Campaign was the real turning point in Europe's wars against France.

The campaigns of 1813–14 and Napoleon's surrender. Napoleon's prospects were not hopeless in the spring of 1813. Opinion in Prussia was still divided about the advisability of declaring war on France. Russia itself had suffered severely during the 1812 campaign, and many influential Russians argued that, their country once made safe, there was no point in fighting to save the rest of Europe. Austrian attitudes were equivocal; Metternich showed no eagerness to join a crusade against France and seems, in some ways, to have preferred the prospect of a negotiated peace leaving Napoleon in power as emperor. (After all, Napoleon was by then married to an Austrian princess—the archduchess Marie-Louise.) Russia and Prussia did, however, conclude the Treaty of Kalisch in February 1813, which was a treaty of alliance against France. During March 1813 the French forces in Prussia were driven back westward, and by mid-March they were back on the line of the Elbe. Meanwhile, Sweden had joined the Allies.

Napoleon's choices, although not entirely clear, were by no means unpromising. His opponents were divided in their political aims, as they later demonstrated at the Congress of Vienna (1814–15). He might have negotiated with them separately, thus dividing them and achieving at least a compromise peace for himself; or he could have

Napoleon's choices in 1813

cut his losses in Spain and restricted his fighting to the area east of the Rhine. But he was unwilling to accept the limitations implicit in such choices, and he chose instead to attempt everything without negotiation.

The campaign of 1813 consisted of two parts. The first part began when Napoleon arrived on the line of the Elbe on April 28. He had by then called up the conscripts of the classes of 1813 and 1814 and had 150,000 men at his disposal in Germany. He wanted to cross the Elbe again, strike at the heart of Prussia before it was fully ready for war, and, by threatening the Russian line of communications, draw both Prussian and Russian armies away from Austria, whose neutrality might thus be maintained. One way to do this was to advance directly on Leipzig and Dresden, compelling his enemies to accept battle or withdraw beyond the Elbe. An outright victory was an essential part of this strategy and was necessary to keep Austria and the states in the Confederation of the Rhine friendly toward France; but that victory was denied to Napoleon. Two inconclusive battles took place—the first at Lützen, southeast of Leipzig, on May 2, in which the Allies were worsted but withdrew in good order; the second at Bautzen on May 20–21, with much the same result. But Napoleon had accepted heavy losses in these battles, and he was at that point much less able to accept them than were the Allies. Further, he signed an armistice with Prussia and Russia at Pläswitz on June 4, which lasted until August 20; this, too, was more helpful to his enemies than to himself—it gave time for Prussia to mobilize its reserves and for Austria to decide to join the Allies—and, when the armistice ended, Napoleon was relatively worse off than when it had begun.

In the second part of the campaign of 1813 the Allies put three armies into the field, comprising some 600,000 men, against whom Napoleon mustered 400,000. The Allies planned to converge from Prussia and Silesia upon Dresden; as on previous occasions, Napoleon's plan was to engage and defeat them separately. A number of distinct but related battles followed. On August 26 and 27 an Austrian army was defeated at Dresden, but on August 30 and then on September 6 the French were defeated, first at Kulm and then at Dennewitz. For Napoleon these defeats were more important than the earlier victory because the Allies continued to threaten him, gaining relatively in strength all the time. During the first part of October the Allies converged on Leipzig, where the Battle of the Nations (October 16–19) took place. By then Bavaria had joined the Allied cause, and Napoleon found himself fighting with an army of 160,000 against combined enemies who could bring against him double that number. Moreover, his enemies knew the danger of accepting battle separately and the need to maintain contact with each other.

Retreat to France

At first Napoleon held his own. But on October 18 the Saxon forces went over to the Allies and on October 19 Napoleon was forced to retreat toward France, defeating a Bavarian force at Hanau on October 30. On November 2 Napoleon's army fell back across the Rhine at Mainz with fewer than 80,000 of the men with whom he had begun the campaign. By then virtually deprived of allies, this campaign was a blow, added to the disaster of 1812, from which Napoleon could hardly hope to recover.

Yet the Allies themselves were still not set upon the overthrow of Napoleon. In November 1813 they offered him peace if France surrendered all its conquests beyond the Rhine, the Alps, and the Pyrenees. As Napoleon knew, however, there were some among the Allies who advocated harsher terms; the British, for example, were anxious to achieve the independence of French-held Dutch and Belgian regions, and that entailed further French concessions. In the end, between Allied differences of view and Napoleon's own doubts about the possibility of continuing in power on such terms, the Allies failed to provide a basis for peace, and they determined to invade France in 1814.

The campaign of 1814, like that of the previous year, was divided into two parts. The Allies planned to move into France from the north via Dutch and Belgian territory, across the Middle Rhine in the area of Mainz and from Switzerland around the Jura Mountains—an overall plan similar to the one they would adopt in 1815. Napoleon, fighting with numerically inferior forces, decided to tackle the Allied armies separately. In the first part of the campaign (late January and February 1814) Napoleon inflicted heavy defeats on the Prussian and Austrian armies advancing from the Middle and Upper Rhine. The Allies appeared to be breaking up at this stage until more determined counsels prevailed. On March 1 Prussia, Russia, Austria, and Great Britain signed a new alliance committing themselves to war and to peace terms only on conditions agreed to by them all.

Napoleon's abdication and exile

The Allies were then set on attacking Paris, and the second stage of the campaign began. Napoleon hoped to draw them off by moving into Lorraine and thus threatening their line of communication, while Paris accepted siege and resisted. The Allies, however, were not diverted by this threat, despite minor local reverses. Moreover, they were by now aware of Napoleon's plan, having captured some of his dispatches; and they also knew that many in Paris were anxious for peace. On March 30 the Allied armies reached the Paris suburbs, and resistance there soon ceased. On April 6 Napoleon abdicated and was soon on his way to exile on the island of Elba, while Louis XVIII returned to Paris as king on May 3.

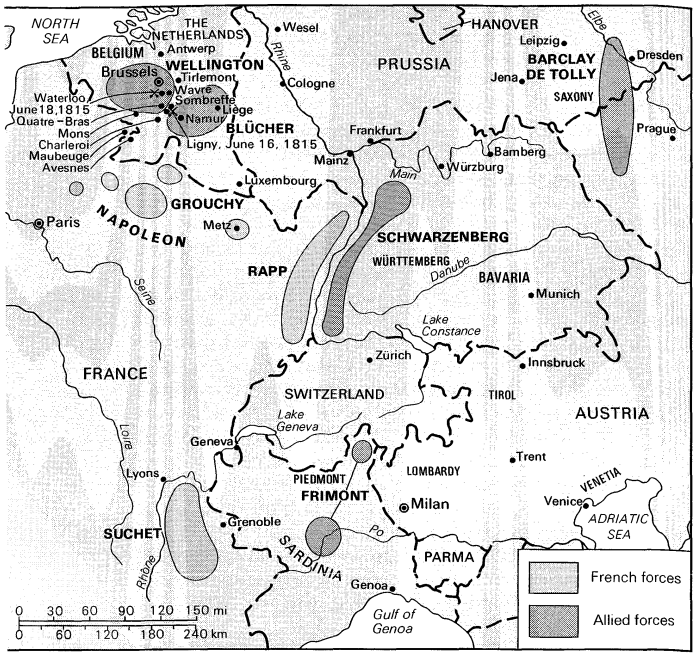
Meanwhile, the Allies settled down to the task of resettling Europe after the long wars; they were still arguing the details when they found themselves at war with Napoleon again.

The Waterloo Campaign. On February 26, 1815, Napoleon escaped from Elba; he landed on the mainland of France near Antibes on March 1. He set out for Paris via the Dauphiné and on March 7 entered Grenoble without any opposition. With towns and garrisons welcoming Napoleon in what proved to be a triumphal journey, Louis XVIII left Paris on March 19; Napoleon entered the capital the next day. A week before that, the representatives of the Allies at Vienna declared Napoleon an outlaw, describing him as "the enemy and disturber of the peace of the world." The three eastern monarchies—Russia, Prussia, and Austria—together with Great Britain, made a new defensive alliance against Napoleon (only a few weeks before, those same powers had been acting more like enemies than friends in their attempts to settle the affairs of Europe), and each undertook to supply 150,000 men for the armies of the alliance and to keep them in the field until Napoleon was finally defeated.

After the failure of some diplomatic manoeuvres designed to detach Britain and Austria from the alliance, Napoleon realized that there was no alternative to fighting; and he prepared himself for an offensive campaign. First, however, he had to raise armies and equip them. There were many veterans of the Prussian and Italian campaigns in France, a large number of them already committed to serving Louis XVIII, and Napoleon found no difficulty in winning them back to his cause. Conscription of new recruits was politically dangerous; and, if Napoleon planned an early offensive, new recruits could hardly be trained in time anyway. So the existing regulars were supplemented by recalling all undischarged soldiers and those who had deserted since April 1814. Thus expanded, Napoleon's army for the 1815 campaign totalled about 250,000, of whom roughly half were available for his striking force and the rest for frontier and garrison duty. In fact, Napoleon did call up the class of 1815 at the end of May, and during June about 50,000 of them were in barracks; but they played no part in the Waterloo Campaign.

The Allied Coalition

The Allied Coalition—the seventh formed against France between 1792 and 1815—planned to raise five armies. An Anglo-Dutch army of 90,000 men commanded by Wellington and a Prussian army of 120,000 commanded by Gebhard Leberecht von Blücher were to operate from the Brussels region into France; an Austrian army of 225,000 men commanded by Karl Philipp zu Schwarzenberg was to operate on the Upper and Middle Rhine; a Russian army of 170,000 men under Barclay de Tolly was to form a strategic reserve; and, finally, an Austro-Italian army of 60,000 commanded by Johann Maria Frimont was to operate from northern Italy. The Allies planned a concentric advance on Paris beginning in late June. Early in April, Wellington left Vienna on his way to Brussels; he



Waterloo Campaign.

Adapted from V. Esposito and J. Elting, *A Military History and Atlas of the Napoleonic Wars* (1964); Praeger Publishers

met Blücher at Tirlemont on May 3. These two did not expect Napoleon to take the offensive against them but agreed that, if he should do so, they would concentrate on the line between Quatre-Bras and Sombreffe. Accordingly, Blücher moved his headquarters from Liège to Namur, and Wellington established his headquarters at Brussels.

Napoleon's plans, as seen above, were basically offensive. He wanted to meet and defeat some of his enemies before the others were ready and then to go on to deal with them piecemeal while they were still concentrating. The enemy armies posing the most immediate threat were those of Wellington and Blücher. Napoleon therefore decided to attack the combined Prussian and Anglo-Dutch forces himself, in command of the Army of the North, 125,000 strong. Leaving generals Jean Rapp and Louis-Gabriel Suchet on the defensive along the Rhine and opposite the approaches from Switzerland and Italy, Napoleon's approach in the theatre he was directly concerned with was in principle similar to his first great campaign in northern Italy in 1796. He faced an enemy stronger than himself but composed of two armies converging on a common line of advance into France. If those two armies could be attacked at their "hinge" (*i.e.*, where their lines of advance converged and joined) and all the more so if they could be attacked separately, they might be prevented from combining and even be driven back upon their divergent lines of communication. Wellington's line ran back through Brussels to Antwerp; Blücher's, via Namur, Liège, and Cologne. Napoleon's aim, therefore, was to advance from Paris to Charleroi and there force himself between the two enemy armies. If they could both be defeated, he would turn back southeastward, join Rapp opposing Schwarzenberg, and then deal with the latter. This strategy was an adventurous approach because Napoleon had few of his old marshals with him, training time had been short, and many of his senior officers were strangers to the men they led.

On June 3 Napoleon ordered the Army of the North to concentrate in the area Maubeuge-Avesnes, about 25 miles (40 kilometres) southwest of Charleroi. On June 14 he moved his own headquarters to the Charleroi area, with Wellington and Blücher still unaware of what the French were doing. By nightfall on June 15, Napoleon's forces were disposed in such a way that they could manoeuvre against either enemy as circumstances dictated. Wellington, in fact, had unintentionally helped—fearing that Napoleon might outflank him on the right and so cut off his line of retreat to Antwerp and the sea, Wellington detached a corps to cover the roads leading from Mons and Ath to Brussels, thus moving his centre of gravity eastward and away from Blücher.

Napoleon's mistakes

Instructing Michel Ney, on the left, to occupy Quatre Bras and thus block Wellington from using the Nivelles–Namur road, Napoleon attacked the Prussians at Ligny, on the Allied left or eastern flank, on June 16. The purpose of this operation was to drive Blücher back eastward and northeastward away from Wellington, and Napoleon counted on support from Ney to reinforce his own left wing to make this possible. The Battle of Ligny was indecisive: the Prussians were, indeed, driven back, but Ney failed to understand Napoleon's strategic plans and did not send the help that might have turned the Prussian retreat into a rout. The next mistake was Napoleon's. He had assumed that, once forced back, Blücher would retreat roughly northeastward via Namur and so move away from Wellington. The Prussian commander, however, was aware of the vital need to maintain contact with Wellington and chose, therefore, to retreat northward in the direction of Wavre. Marshal Emmanuel de Grouchy, who had been instructed to pursue the Prussians, thus took the wrong road; and Napoleon realized too late that his assumption was incorrect.

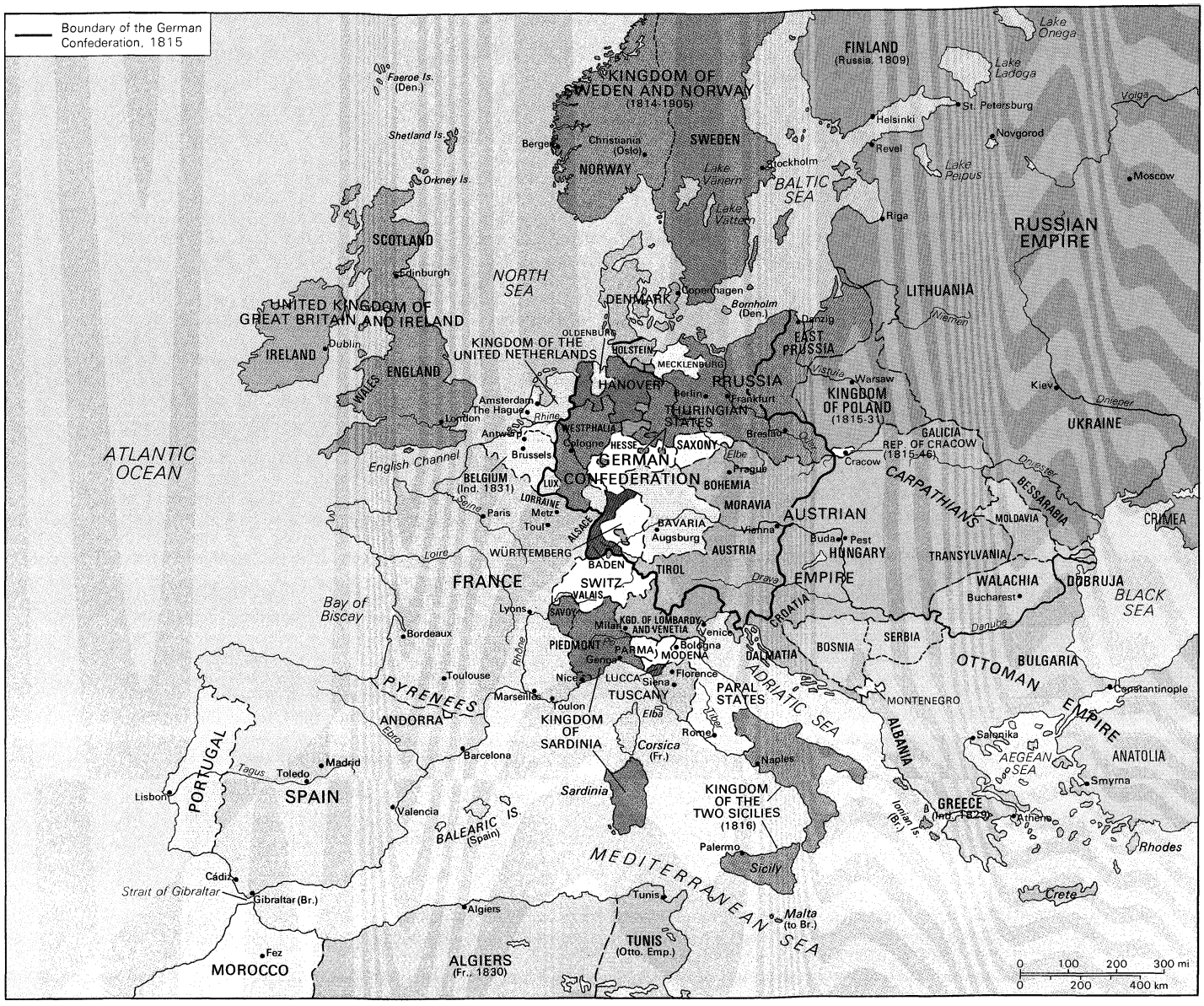
The third mistake was also Napoleon's. The essence of his earlier strategy in such a situation was to act more quickly than his enemies and, by gaining even a few hours, to deny them time to settle down and recover. This time Napoleon's conduct of operations fell below his own best standards. The night of June 16 and the morning of June 17 were wasted. When at last Napoleon began to move toward his left at about noon on June 17—a move designed to attack and defeat Wellington—the latter was already well aware of Napoleon's plans and had been given time to draw his army up in a strong defensive position just south of Waterloo. Moreover, Napoleon's march, already begun late, was further hampered by heavy rain. As a result, when the Battle of Waterloo was fought on June 18 all element of surprise had been lost. And by the early evening on the 17th Blücher kept his promise to Wellington and appeared on the battlefield with invaluable reinforcements, turning the scales against Napoleon at the critical time. Waterloo was Napoleon's last battle.

Up to the morning of June 16 Napoleon's preparations for the Ligny–Waterloo Campaign suggested that his military genius was unimpaired. He moved faster than his opponents, and, as far as it lay in his power, he laid the foundations of victory. It was in his conduct of operations from the evening of the 16th until midday on the 17th that he failed.

After Waterloo, Napoleon retreated southward, and on June 21 he entered Paris. The next day he abdicated. A fortnight later Wellington and Blücher also entered Paris, bringing Louis XVIII back with them; 23 years of war were ended. (N.H.G.)

The peoples and states of Europe that had experienced the events of the French Revolutionary and Napoleonic era, however, would never be the same again. In 1806 the Holy Roman Empire had been destroyed, never to be restored. The multitude of little German states, pounded by Napoleon's conquests, grouped into new formations. Prussia had undergone drastic reorganization and emerged with new prestige and strength. Through the legal codes and administrative systems that Napoleon had introduced into conquered territories, new ideas and standards of government had become familiar and popular. Even the revolutionary ideals of popular sovereignty and national rights had been spread by his conquests, and his harsh rule had stirred a new spirit of national solidarity and resentment.

Napoleon's despotism was in many respects paternalist and reminiscent of the prerevolutionary benevolent despots. He patronized the arts and encouraged scientific research. He carried out schemes of public works and rationalized civil administration. In other ways he foreshadowed the tyrants of the 20th century, for he governed through a police state, censorship, manipulated elections, and plebiscites. The extent of his conquests was due in part to his genius for war and diplomacy but in part



Europe after the Congress of Vienna.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

also to the selfish quarrelsomeness and ambitions of his adversaries. Prussia sought gains at the expense of Austria, Russia at the expense of Turkey and Poland, while Great Britain expanded its commercial empire and its sea power. Only experience of repeated separate defeats forced the powers into coalition against him. But in the end he was defeated by the professonal armies of the major powers. It was the old-time dynastic monarchs of Europe who assembled as the victors to reconstruct a shattered Europe. Yet it was a new Europe, seething with new ideas and beset with new needs, that confronted the peacemakers of Vienna in 1815.

The Restoration. A quarter-century of revolutionary ferment and recurrent warfare not only made reconstruction and restoration very difficult: it foredoomed to failure any such labours that did not recognize the extent of the transformation these upheavals had brought about. The generation after Waterloo was one of peace between the powers; to that extent the Vienna settlement succeeded. It was also one of restlessness and revolt within most states and of deep social conflicts; this was the irrepressible aftermath of the events of 1789–1815. These internal disturbances in turn affected relations between the powers and brought about the downfall of the "congress system" by which the peacemakers hoped to preserve the peace. Though there was no war between the major powers until the Crimean War, peace was bought at the price first of general repression and the subordination to Austria of the German Confederation, then of periodic crises and realignments of the powers.

The Congress of Vienna

The Treaty of Paris of November 1815 pushed back French power to the frontiers of 1790 and exacted a large indemnity. The Treaty of Vienna of June 1815 was agreed to, in effect, by the representaives of the five major powers at the Congress of Vienna: the tsar Alexander I of Russia; Metternich, the chancellor of Austria; Karl August von Hardenberg, acting for the king of Prussia; Lord Castlereagh for Great Britain; and Talleyrand, the wily spokesman of France. In the general Congress nearly every European state was represented, and its purpose was a general settlement of European affairs.

The guiding principles were to prevent a recurrence of French aggression, to restore territories as far as possible to their legitimist rulers, and to establish a balance of power between the major states. To these ends territories were allocated among the powers regardless of the wishes of their inhabitants. Belgium and Holland were combined into the Kingdom of The Netherlands, as a buffer state against France. Norway was transferred from Denmark to Sweden. The Palatinate on the left bank of the Rhine went to Bavaria. The German provinces of Saxony, Westphalia, and the Rhineland as well as the Polish provinces of Poz-

nan (Posen) and Pomorze and the remainder of Swedish Pomerania, went to Prussia. Germany was formed into a German Confederation of states dominated by Austria, which also took Lombardy and Venetia. Poland again disappeared from the map as an independent state. Russia and Great Britain kept their conquests: Britain thus gaining Malta and Heligoland, the Dutch colonies at the Cape, Ceylon, and Mauritius, Russia obtaining Finland, the major part of Poland, and Bessarabia.

The spoils having been divided on a nice calculation of mutual gains and a judicious balance of power, the settlement was to be preserved by periodic meetings of the major powers to discuss any unsettling issues that might arise. By the terms of the Quadruple Alliance, Great Britain, Prussia, Austria, and Russia pledged themselves to maintain, by force, the terms of the settlement for a period of 20 years. But the concert of Europe, to be conducted at periodic congresses, soon broke down. Britain differed fundamentally from its allies about their readiness to intervene in the internal affairs of other states (even of France) in order to keep the settlement intact. At the Congress of Aix-la-Chapelle in 1818 and again at the Congress of Troppau in 1820, Britain refused to take part in the policing of Europe. This main task fell on Austria, which accepted it willingly enough, and Metternich established his famous system of espionage and repression for dealing with any signs of agitation or insurrection.

The greatest weakness of the Restoration was that it was wedded to dynastic principles at a time when all the main forces of change were opposed to dynasticism. In France Louis XVIII issued a charter of rights and set up a limited parliamentary regime. Patently owing his restoration to the allies and arousing more apprehension than loyalty in his subjects, he survived by treading warily and restraining the extremists. Other restored monarchs, less shrewd than he, such as Ferdinand VII of Spain and his uncle Ferdinand I of Naples and Sicily, stored up trouble for themselves or their successors by policies of vindictiveness and blind reaction. Napoleon, by his own upstart dynasticism, which led him to put members of his own family on the thrones of conquered countries and to marry, in 1810, a Habsburg archduchess, had done as much as the Revolutionaries to discredit dynastic monarchy and destroy its mystique.

Revival of church power

The Restoration brought back the church as well as the monarchy, for the alliance between altar and throne was still traditional in Europe. The Roman Catholic Church regained a highly privileged position; and, where, as in France, it could not recover all its lands and property, it was generously supported by the state and resumed its control over education. It reaped some of the ultimate advantages of having been persecuted and of enjoying the prevalent reaction against the rationalism and free thinking of the 18th century. Faith revived, as did clericalist sympathies; though soon Ultramontane claims met resistance even from the restored monarchs.

Liberalism and nationalism. Europe of the Restoration was a collection of old bottles filled with new wine, and trouble was soon fermenting. In Spain, Portugal, Piedmont, and Naples in 1820–21 revolts broke out, demanding democratic constitutions, the model for which was the abortive Spanish constitution of 1812. Austrian troops quelled the uprisings in Piedmont and Naples, but those in Spain and Portugal temporarily succeeded, because Great Britain and France refused to join in repressive intervention. In 1821 the Greeks rebelled against their Turkish rulers, raising for the powers the greater peril that Russia would take the chance to attack Turkey in the popular cause of a Christian nationalist movement against infidel misrule. The Spanish and Greek risings were the chief concern of the Congress of Verona in 1822. In 1823 France invaded Spain and restored King Ferdinand VII.

Austrian domination was confirmed by the suppression of revolution in Italy and Spain, but Great Britain emerged as a champion of movements for national independence and constitutional government in Greece and Portugal. Both the Quadruple Alliance and the congress system were disrupted by these events. The Greeks fought on, rousing Philhellenic enthusiasm throughout Europe by their heroism; and, when they won independence in 1830, it was because Britain, Russia, and France went to war with Turkey and forced it to make concessions in the Treaty of Edirne (September 1829). In 1826–27 Britain had also to intervene in Portugal to preserve the forms of constitutional government.

Disruption of the new system paved the way for wider movements of insurrection in 1830. France raised the flag of revolution in protest against the Ultramontane and aristocratic reaction of Charles X, who had succeeded his more tactful brother in 1824. The liberal opposition could take its stand on the charter of June 1814. In July 1830 republican groups in Paris seized the Hôtel de Ville and raised the tricolour flag of the Revolution. By a manoeuvre of the moderate liberal politicians, journalists, and bankers, Charles, who fled to England, was replaced not by a republic but by Louis-Philippe, representative of the Orleanist branch of the royal family, a more accommodating, liberal-minded man than his immediate predecessors on the throne.

Britain's Parliament gained greater control over government and rested on a wider (though still restricted) franchise. The new regime closely resembled the limited constitutional monarchy of Great Britain after the Reform Act of 1832 widened the electorate.

In 1830, too, in addition to the Greeks' attainment of independence, the Belgians won their independence from the Dutch (so destroying one provision of the Vienna settlement). Ripples of revolution, stirred by these events, spread even to lands under the watchful eye and the heavy hand of Metternich. In Brunswick, Hanover, Saxony, and other states of the German confederation, local revolts forced concessions from reluctant rulers. In such states as Baden and Bavaria, which already had parliamentary systems, liberal oppositions gained in elections and the press became more openly critical of their governments. In 1832 Prussia and Austria got the Diet of the German Confederation at Frankfurt to pass the "six acts," curbing press and opposition in all German states. By 1835 Metternich's system was again secure in Germany. In Italy, too, it faced revolts in Modena and Parma and in papal territories, but these Austria quickly quelled. A Polish insurrection was similarly crushed by Russia in 1831. Only in free Switzerland did most of the cantons succeed in setting up more liberal constitutions.

By the 1830s domestic conflicts between conservative, dynastic, and clericalist forces on one side and liberal, nationalist, and republican movements on the other had become a general pattern in European politics. Internationally, this same division was reflected in the realignment of Great Britain, France, and Belgium as against Austria, Prussia, and Russia. It was a pattern very different from anything envisaged in 1815. Most liberals (seeking parliamentary and more representative constitutions) were also nationalists (seeking the unification and independence of all communities feeling themselves to be distinct nationalities). The ideal of popular sovereignty, born of the French Revolution, implied both internal and external self-government. Government and governed were to be brought into closer reciprocal relations, in the name both of democracy and of national self-determination.

International realignment

A link between domestic forces of revolution in different lands was formed by the secret societies, which flourished because the police states of the Restoration drove all rebels undergound. It was a golden age of insurrectionary conspiracies and feverish plots, many of them wild and impracticably idealistic, though the Carbonari in Italy and Spain were at times both active and effective. Metternich's spies were equally active internationally against the secret societies; and, if he sometimes exaggerated the danger, it was to justify his own penetration of other regimes. In such societies cultured aristocrats, intellectuals, and poets mingled strangely with cutthroats and adventurers, many of them spending a large portion of their lives in prison or exile.

This sensational quality of the politics of the time matched the mood of the Romantic Movement in culture, though it had close affinities with both liberalism and nationalism (see below). The nationalist poets of France, Poland, and Greece were romantics, hostile to the Restoration as well

as to Napoleon because they exulted in individual freedom and personal heroism. They dreamed of a society of free and equal individuals as well as of free and independent peoples. Many, like the French poet Lamartine, who became minister of foreign affairs in the Second French Republic of 1848, assumed leading roles in politics. The prototype of the romantic liberal patriot was Giuseppe Mazzini, a member of the Carbonari and founder of the Young Italy movement for Italian unification under a democratic republic. Such men played a central role in the events of the great year of revolutions—1848.

The spread of revolt

In 1848, as in 1830, discontent grew from bad economic conditions as well as from liberal and nationalist aspirations. The flames of revolt spread fast from one city to another—for it was mainly a revolt of the towns, where social distress was acute and where organized insurrection could more readily occur. In January 1848 the King of Naples again faced open rebellion, and within a month riots took place in nearly all the large Italian cities. In February Paris put up barricades and overthrew Louis-Philippe, proclaiming a Second Republic. Within the year both types of revolt—national revolt against foreign rule and the dreary repressiveness of Metternich's agents and democratic revolt against the inadequacies of middle-class government—were repeated elsewhere. Disturbances in Great Britain (where the Chartist movement for radical political reforms had been vainly active for more than a decade) and in Belgium and Switzerland were on the French pattern. The movement failed in Great Britain. It succeeded in France only until 1851, when Louis-Napoléon, nephew of the great Napoleon, seized power. Switzerland alone achieved the lasting result of a new federal constitution (1848). The importance of the revolutions lay in the outburst of national enthusiasm, which spelled the end of Metternich's systems in Europe, and in the schisms that they caused between liberalism and nationalism.

The Italian risings revealed throughout the whole peninsula enthusiasm for national independence but deep differences as regards the means to this end. A party of conservative liberals pinned its hopes on Charles Albert, king of Piedmont and Sardinia; a party of Catholic federalists looked to leadership from the papacy; yet a third party, composed of Mazzinian republicans, fought for a democratic republic. Weakened by these divisions, the revolution was eventually crushed by Austria. Within the nationally diversified empire of the Habsburgs the risings were violent enough in Vienna to force the resignation of Metternich himself and to exact the promise of a new constitution; while in Hungary a new national leader appeared in Lajos Kossuth demanding home rule for the Hungarian population at the expense of Slav minorities within the kingdom. Here, too, deep divisions brought disaster to the nationalist cause, and by 1849 the Habsburgs were back in power, ruling through the army and the bureaucracy.

In Germany a body representing the aspirations of all nationally minded Germans met at Frankfurt in May 1848 and sat for a year. But irreconcilable divisions appeared between the "great Germans," who wanted a federal state including all Austrian lands except Hungary, and the "little Germans," who looked to Prussia for leadership and wished to omit the mixed peoples of Austria in order to unify Germany tightly. Roman Catholics tended to favour Austrian leadership, Protestants to favour Prussian. By 1849 the King of Prussia had withdrawn concessions made within his own kingdom and refused the crown of a united Germany. Liberal nationalism was discredited throughout Germany.

The paradoxical outcome of the year of revolutions was military rule in Prussia and Austria, a restored confederation in Germany, and a revived Bonapartism in France, But though frustrated in 1848 and subjected to even more severe repression by 1850, nationalists everywhere now had a clearer vision of how their aims could be reached. Governments, having tasted the strength of nationalist feeling, were forced to consider how they could harness it instead of merely crushing it. The system of Metternich had gone and seemed even more old-fashioned than the heroics of Romantic poets and secret societies. A mood of grimmer realism pervaded European affairs after 1850.

In 1850, as in 1815, it was the professional armies of monarchical governments that won; the intellectuals and romantic idealists went down in defeat. Again, strong monarchies seemed reaffirmed in their lease of power; and the power of the Roman Catholic Church revived, and its role in education grew. Yet a certain residue of more liberal government remained. Universal suffrage and parliamentary institutions were not abolished in France, though the emperor Napoleon III (as Louis-Napoléon became) blunted their edge. In Prussia, as in Sardinia, parliamentary institutions remained. In Austria seigniorial rights were not restored, and the new rulers of these countries differed in temperament and methods from the traditionalist monarchies of the restoration. Above all, the forces of profound economic change were now at work throughout the whole of Europe, making further changes in politics and in mental attitude inevitable.

New national states in the Balkans. From the end of the 18th century the Turkish power was crumbling in southeastern Europe, and the rival ambitions among the powers increased. Great Britain and France were interested in Egypt and the Levant, Austria and Russia in the Middle East and the Balkans. These problems entered a new phase with the stirring of liberal and nationalist movements in eastern Europe, especially in Poland and Greece. Greek independence involved joint action by Britain, France, and Russia against Turkey. When, in December 1832, an internal crisis in Turkey prompted the Sultan to appeal to Russia for help, the tsar Nicholas I sent an army and a fleet to defend Constantinople. In 1833 Sultan and Tsar signed a treaty closing the straits to other powers and creating a masked Russian protectorate. Seven years later Great Britain joined with Russia, Prussia, and Austria to intervene and make a convention (1841) with the Sultan whereby he undertook to close the straits to all warships. Each power showed its willingness to interfere in Turkish affairs when it felt that its own interests might be affected, and each feared the machinations of the others in this turbulent area.

The Crimean War

It was these entangled interests that led, in 1854, to the first war between the former allies of 1815—the Crimean War—in which Britain, France, and Sardinia, allied to Turkey, made war on Russia. The Russians, defeated by the Turks on the Danube, evacuated the Romanian principalities that they had occupied, but the war continued in the Crimea until the fall of Sevastopol. Austria, as a nonbelligerent, discussed with the allies conditions of peace to be imposed on Russia. In 1855, when the tsar Alexander II succeeded his father, Russia agreed to the terms proposed; and at the Congress of Paris (1856) the great powers, with Sardinia and the Sultan, made a peace treaty with Russia. The Sultan promised reforms in return for a guarantee by the powers of the integrity of his empire. The Black Sea was neutralized and closed to ships of war. The Romanian principalities were declared autonomous, and Russia restored part of Bessarabia to Moldavia. The Congress was a revival of the idea of a concert of Europe, and its declaration respecting maritime law revived concern for a public law of Europe. None of the settlement was to prove durable. The chief beneficiaries of the war were Napoleon III and the King of Sardinia, both of whom gained in prestige. The Eastern Question was not settled.

This became apparent in 1876, when the Slav Christian subjects of Turkey in Serbia and Bulgaria again rebelled, and their savage suppression by Turkey attracted the attention of the powers. Russia went to war, and in the Treaty of San Stefano (1878) arranged for the formation of an enlarged Bulgaria under Russian protection. Again the concert of Europe revived at the Congress of Berlin, attended by all the great powers, including the new Italy and the new Germany, the only two powers that left it without some territorial gains. The settlement, however, left each power dissatisfied and more anxious than before.

The remaking of central Europe, 1850–71. The schism between the forces of liberalism and nationalism, which appeared by 1850, and the breaking of the spell of peace between the powers by the Crimean War inaugurated a new era in great-power relationships. To forge nation-states and win their independence became the task of

strong leaders of existing states, with armies and economic resources of their own, not the dreams of exiles or conspirators. Diplomacy and war were the new agencies of successful nationalism.

Economic changes reinforced this tendency. Railways made it possible to transport men and goods—including armies and war materials—rapidly from one part of Europe to almost any other. The expansion of industrial productivity opened the way for the rapid growth of the economy of any state ready to take advantage of these new opportunites. France under Napoleon III was well placed to seize these opportunities, and the Second Empire (1852–70) was an era of growth, prosperity, and expansion under a paternalist, if harsh, regime. Exports exceeded imports, so that France, like Great Britain, could invest capital abroad. The Comité des Forges, heading France's great iron industry, was formed in 1864 and extended its interests into Belgium and Germany. A new French colonial empire was built up. Algeria was completely taken over by 1857; Tahiti and the Ivory Coast had been added earlier. Expeditions went to Peking in 1859–60 and to Syria in 1861. Explorers went to West Africa, and new settlements were made in Dahomey and the Guinea coast. New Caledonia was occupied in 1853. From 1859 onward three provinces were annexed in Cochin China, and a protectorate was set up over Cambodia (1863). At home the Emperor perpetuated Bonapartist traditions by his schemes of public works, notably the replanning of Paris. France in those years became, like Great Britain, a world colonial power and a larger, richer, more powerful state.

French colonial empire

In central Europe two states were advantageously placed to seize the new opportunities of growth. One was the constitutional monarchy of Sardinia, Piedmont, and Savoy, with its favourable strategic position in northern Italy, its enlightened government anxious to westernize the country, and its prestige as a leader of Italian hopes of national unification. The other was the kingdom of Prussia, already advanced in internal administration and in fiscal and military organization. With territories and interests widespread throughout Germany, it was the natural focus for all hopes of national unity excluding Austria. Austria of the Habsburgs was precluded from being a champion of nationalist aspirations anywhere, because its sprawling domains included peoples of many races and national loyalties and because its survival as an empire depended on discouraging and suppressing movements of emancipation.

In 1852 Count Camillo di Cavour became premier of Piedmont; in 1862 Otto von Bismarck became prime minister of Prussia. The reshaping of central Europe, achieved by 1871, was preeminently the work of these two states and these two men.

Cavour, as Piedmont's minister of agriculture, commerce, and marine in 1850 and soon also as minister of finance and premier, developed intensively his country's whole economy. He built or improved railways, roads, ports, docks; he expanded trade and linked Piedmont to the West by loans and commercial treaties. He spent money generously on the army, especially after his successes in the Crimean War. He agreed with Napoleon III in 1858 to make war jointly on Austria and then provoked Austria into war (1859). After six weeks of fighting the Austrians were driven from Lombardy, and Napoleon abruptly made terms with Austria, on his own, at Villafranca. But he had precipitated the first triumph of Italian nationalism. Cavour gained Lombardy but not, as he had hoped, Venetia. In 1860 Parma, Modena, Tuscany, and Romagna adhered to his cause, while Giuseppe Garibaldi conquered Sicily and Naples, and Cavour annexed all the central Italian states except Rome itself. When the first Italian parliament met at Turin in January 1861, the new kingdom included all the Italian peninsula except Venetia and the papal city of Rome. Cavour died in June 1861. The business that he left unfinished was completed by 1870 under the momentum he had given: Venetia was added in 1866 when Prussia defeated Austria; and in 1870 Germany's defeat of France left open the road to Rome.

Bismarck, as a loyal servant of Prussia, regarded it as his duty to assert Prussian as against Austrian leadership of Germany. To this end he sought alliances with France and Russia. Prussia was growing in population and in power. Since 1815 its population had risen from 11,000,000 to 18,000,000. Its economy had expanded; and, like Cavour, Bismarck was convinced of the need to spend lavishly on armaments and the army. He despised parliamentary assemblies and liberal ideas and believed in the traditional Junker (Squire) values of order, discipline, service, and duty. But his task was formidable. Since 1815 Germany had been divided into 38 states, unequal in size and diverse in constitution, bound together loosely within the German Confederation (Bund), which included Austria as well as Prussia.

Bismarck's first task was to drive Austria from the Bund. First Prussia joined Austria in defeating Denmark in 1864. As a result (Convention of Gastein, 1865) they took control of the duchies of Schleswig and Holstein, Prussia being responsible for the administration of Schleswig, Austria for that of Holstein. The future of the duchies remained the joint concern of Austria and Prussia—a convenient bone of contention for future use. Then Bismarck isolated Austria diplomatically by promising Italy Venetia in return for Italian help against Austria and by ensuring French neutrality. In June 1866 he forced the issue by proposing an end of the Bund and the election of a German assembly to draft a new consitution excluding Austria. After defeating Austria at Königgrätz (Sadowa) on July 2, he made the Treaty of Prague in August, annexing both Schleswig and Holstein and setting up a North German Confederation under Prussian leadership. The southern German states formed a looser association "with an independent international existence." Austria undertook to claim no further part in German affairs and surrendered Venetia to Napoleon to hand on to Italy. The efficiency of the Prussian Army, the precision of the campaign, and the skill of Bismarck's diplomacy achieved no less (and no more) than Bismarck intended. The war was a sharp blow to French prestige. It brought Italian and German unity one step nearer. It helped to bring about reorganization of the Habsburg Empire into the dual monarchy of Austria-Hungary in 1867.

The process of German unification was completed by the Franco-German War of 1870. Napoleon III—his government at home weakened by growing liberal and republican opposition, his prestige shattered by the failure of his Mexican expedition and by the demonstration of Prussian power at Königgrätz sought feverishly but fumblingly for some form of compensation. The conventions of diplomacy and the concepts of a balance of power required France to seek some material gains to offset the recent accessions of power to Prussia and Italy. Bismarck used these anxieties to precipitate war with France—a war in which France appeared as the aggressor. The pretext for conflict was the candidacy of Leopold of Hohenzollern for the vacant throne of Spain; the possibility of a German prince on the Spanish throne infuriated France. But even when Leopold withdrew, the war parties of Paris and Berlin were so intransigent, the public feelings of the two countries so frantic, that France declared war in July. At the Battle of Sedan the main French Army and even Napoleon III himself were captured; and the war would have ended then had not a new republican government of national defense in France continued hostilities until February 1871.

The Battle of Sedan

The war achieved Bismarck's aim. In January 1871 King William I of Prussia was proclaimed German emperor at Versailles. Germany was united under Prussian hegemony when the southern states combined with the North German Confederation within a new imperial constitution. France, after an interlude of civil war in the Paris Commune of 1871, set up the Third Republic in 1875—a parliamentary regime of a moderate, conservative kind. Russia took the chance to repudiate the Black Sea clauses of the Treaty of Paris of 1856, and the other powers were too much preoccupied to do more than protest. A completely new balance of power existed in Europe, with Germany overshadowing France and Russia overshadowing Austria.

Rise of Socialism. The history of European Socialism,

like that of liberalism and of nationalism, entered a new phase around 1850. Before midcentury, Socialist ideals derived partly from the democratic ideals of the French Revolution and partly from a wide variety of humanitarian, philanthropic, and idealistic trends of thought and action. The leading Socialist thinkers had been men like Robert Owen in England and Charles Fourier in France, stirred by the social injustices and individual hardships of the Industrial Revolution (see below) with its exploitation of unorganized labour. They looked to enlightened employers, public conscience, or new modes of economic and political organization to improve the material lot of the working classes. They went beyond this, too, and saw in a transformation of social ideas and habits the only path to a more just way of life for all. They pinned their faith on free association and education to bring progress. They wanted a new kind of society.

By midcentury, Socialist thought, like nationalist thought, was moving toward greater reliance on state action and legislation. Louis Blanc in France and radical reformers in Great Britain came to urge more positive action by governments to protect workers in factories. They sought to implement the "right to work" by policies of poor relief and provision of employment in national workshops. These more empirical and realistic proposals were in sharp contrast with two other strands of Socialism: the idealist and visionary Socialists, who encouraged migration to the American frontier to form simple egalitarian communities practicing communism in primitive agrarian conditions, and the Socialists of the secret societies and underground conspiracies, who plotted mass uprisings in the capitals of Europe to overthrow existing regimes and establish extreme democratic republics based on universal suffrage.

All alike suffered severe reverses in the debacle of 1848–49. The insurrectionary followers of Auguste Blanqui in France and the international supporters of Filippo Buonarroti in France, Italy, and Belgium attempted risings that were brutally crushed. Louis Blanc, who served for a time in the provisional government of the Second French Republic of 1848, found his aims frustrated and discredited. The only successful Socialist movements by midcentury were the moderate trade-union movements and the cooperative movement promoted in Great Britain by Robert Owen. They were not revolutionary and not even political in purpose. The Industrial Revolution had not yet produced the social conditions needed for a broad and effective Socialist movement anywhere. In every European country approximation to universal male suffrage gave power to the peasants, who proved, by their support of reaction in France or in Hungary in 1849, to be a predominantly conservative force. But the movement from countryside to town continued everywhere, and the existence of large industrial urban populations would change the whole situation.

The Communist Manifesto

In 1848, on the eve of the ill-fated revolutions, Karl Marx and Friedrich Engels issued *The Communist Manifesto.* It attacked all previous Socialist movements as "utopian" and based its program of revolution on an analysis of the inherent tendencies in capitalist society. Finding in private ownership of the means of production the source of inevitable class war between those who own the means of production (bourgeoisie) and those who have only their labour to sell (proletariat), they saw in the growing class of industrial workers and landless peasants the truly revolutionary class of the future. Only when this class seized power, liquidated the bourgeoisie and instituted a dictatorship of the proletariat could modern society move forward to a completely classless, Communist society.

Their famous pamphlet had no effect in 1848. But in the aftermath of disillusionment during the 1850s their arguments found readier adherents. An industrial exhibition in London in 1862 brought together British and French working-class leaders, and by 1864 the International Working Men's Association was formed by delegates from Great Britain, France, Germany, Poland, Italy, and other countries. Marx drafted its rules; it organized periodic conferences and became the basis of the First International. Marx deepened and strengthened his analysis in his study of *Das Kapital* (of which the first volume appeared in 1867). He began to attract disciples in various countries, especially in his native Germany.

The events of 1870–71 had both adverse and beneficial effects on the fortunes of Socialism and Communism. The Paris Commune of 1871 was not essentially Marxist in its inspiration, though some Marxists took part in it. Its defeat was a grave reverse for all Socialist and working-class movements in France. Marx depicted it (in *The Civil War in France*) as a prototype of the new proletarian revolutonary technique and hailed it as a sign of hope. In fact, it was the last flicker of the old Jacobin tradition of the barricades, and by leading to the killing or exiling of Socialist leaders and discrediting revolution it secured a highly conservative regime in France for the next 70 years. On the other hand, the creation of unified states in Italy and Germany, each with parliamentary institutions within which Socialist parties could operate, paved the way for the growth of large and powerful Social Democratic parties during the next generation. (D.Tn.)

Events outside Europe. Outside of Europe some developments of considerable significance, especially for the longer term, also took place. The long self-imposed isolation of Japan came to an end when, in 1854, the American commodore Perry succeeded in negotiating the Treaty of Kanagawa. After a period of hesitation, in 1868 Japan deliberately embarked upon the path of Western imitation as the only manner in which it could retain its independence.

The opening of Japan

In China, a process of decline and disintegration continued reminiscent of that of the Ottoman Empire. The Taiping Rebellion eventually led to foreign intervention, mainly Anglo-French, the forcible opening of China, and the imposition upon that country of discriminatory treaties. The Russians were expanding in Siberia, securing the boundary of the Amur River. Vladivostok, on the Pacific, was founded in 1860.

In Southeast Asia the British influence was reaching out from India, reorganized after the great Indian Mutiny of 1857, toward Burma and Siam, while the French established themselves in Cochin China and Cambodia, thinking of the Mekong River as a possible southern route of access to China.

The American continent was supposedly out of bounds for European intervention, but the British and French maintained interests in Mexico and in Texas; Britain's maritime and commercial empire made the possibility of an isthmian canal of equal interest to Britain and to the United States. Manifest destiny meant territorial expansion for the latter, mainly at the expense of Mexico, but domestic issues exploded into civil war between the North and the South in 1861. Aware of the potential rivalry of the United States, the British and the French, at the governmental level at least, tended to be sympathetic to the fragmentation of U.S. power and toyed with recognition of the South. The war gave rise to some incidents—the "Trent" affair and the "Alabama" claims—but the ultimate success of the North preserved the unity of the country while laying the bases for its subsequent astounding growth. The emergence of the United States, Germany, and Japan within less than a decade is worth noting.

The French in Mexico

The U.S. Civil War facilitated French intervention in Mexico, and Napoleon III succeeded in placing his candidate, Archduke Maximilian of Austria, on the Mexican throne in 1863—an action highly unwelcome in the United States. The termination of the Civil War and the menace of a united Germany induced Napoleon III to recall the French force from Mexico, and, lacking local support, luckless Maximilian was captured and executed. The Mexican episode contributed to the impression of the French emperor as a meddler, a factor that, in turn, contributed to British and American indifference during the Franco-German War of 1870–71. (Ed.)

THE INDUSTRIAL REVOLUTION

From the mid-18th century until World War I, economic history centres around a group of changes known as the Industrial Revolution. The term industrial revolution is commonly used to denote those changes in the processes and organization of production that mark the passage from

an agrarian, handicraft economy to one dominated by industry and machine manufacture, from what is sometimes called a premodern or traditional economy to a modern one. The term, when capitalized, refers specifically to the first historical instance of this transformation, which began in Britain in the 18th century, spread from there to the Continent and to offshoots of Europe overseas (the United States, in particular), and transformed in the span of a century the life of Western man, the nature of his society, and his relationship to the other peoples of the world.

At the heart of the Industrial Revolution was a cluster of innovations in the technique and mode of industrial production: (1) the substitution of inanimate for animal sources of power (in particular, the introduction of steam power fuelled by coal), (2) the substitution of machines for human skills and strength, (3) the invention of new methods for transforming matter (in particular, new ways of making iron and steel and industrial chemicals), and (4) the organization of work in large, centrally powered units (factories, forges, mills) that made possible the immediate supervision of the production process and a more efficient division of labour. These innovations in industry promoted and were, in turn, supported by major changes in the technology of agriculture and transport.

Underlying all of this was the systematic application of knowledge to the devising of more efficient production. In the first part of the Industrial Revolution, the knowledge was mostly the result of practical experience and informed empiricism. With time, however, beginning in such industries as chemical manufacture, applications were derived from pure science, and, by the end of the Industrial Revolution, the relationship was reversed. Theory had moved ahead of practice in most domains, and inventors were tapping the pool of scientific knowledge for usable ideas and information.

The Industrial Revolution was essentially completed in the first industrializing countries by the late 19th century. By then they had made their passage to the new technology and were pursuing new paths of change. Again, there was a clustering of innovations that gave the impetus to this further growth: electrical power, the internal-combustion engine, petroleum fuel, the automobile, and science-based chemical manufacture (the first synthetics). Some have called this array of changes a second Industrial Revolution. In the second half of the 20th century, a similar clustering of innovations (*e.g.*, atomic power, electronics, and computers), is effecting perhaps a third Industrial Revolution. Each of these stages has been marked by important gains in the production and delivery of energy; in the speed, precision, and convenience of tools and machines; and in the effectiveness and usability of final products (standardization of parts, refinements of tolerances, miniaturization of components).

Changes in productivity

All of these technological advances translated into economic growth—that is, increased productivity and income per head. Productivity is measured by the ratio of output to one or more factors of input (labour, capital, materials). As a result of the Industrial Revolution, labour productivity increased thousands of times in some branches of industry (cotton spinning, for example), hundreds of times in others (weaving, steelmaking, chemicals), with corresponding decreases in the cost of manufacture. By contrast, those activities little affected by the new techniques—skilled arts and crafts, for example, or barbering—showed almost no gain in productivity and little reduction in costs. The effect was a massive shift of demand in the direction of those products and services the costs of which fell the most and an increase in the numbers of persons and amount of capital employed in the branches furnishing these products and services. This shift was reinforced by the gains in income per head, which made it possible for the populations of advanced industrial countries to devote a smaller portion of their income to necessities, and to spend more of it on luxuries or "extras." This expenditure over and above the requirements of subsistence is especially responsive to the changes in relative prices entailed by changes in relative costs of production. The effect was a substantial growth of those branches most affected by the new technology and, in general, a shift of resources (including labour) out of agriculture into industry and services and a movement of population from the countryside into the cities and towns.

The Industrial Revolution in Britain. The first country to make the transition to modern industry and economic growth was Britain. The reasons for this precocity were partly material, partly social and institutional. Britain was favoured by its oceanic position, which gave the country easy access to markets and suppliers overseas; by its modest size and highly indented coastline, which placed most of the country within easy reach of water transport; and by the abundance of those natural resources that proved to be particularly important for the new industrial technology. These included water (for waterpower, chemical processes, textile finishing, and, in the form of high humidity, for the spinning of yarn); salt, for the manufacture of industrial alkalis and acids; china clay; iron and nonferrous ores; and, above all, huge, accessible deposits of coal, for use as fuel and as a source of carbon in the smelting of iron. The one resource Britain lacked was timber, but this very handicap was turned to an advantage by the precocious recourse to fossil fuel.

At the beginning of the 18th century, Britain was a medium-sized nation. Its population of nearly 6,000,000 people was less than one-third that of France; yet, Britain constituted the biggest single market in the world, not only because of the island's compactness but also because the movement of goods through the country was unimpeded by man-made obstacles. France, Britain's chief political and economic rival, was divided into three major customs areas, and traffic was further trammelled by a multiplicity of special tolls and by customary restraints on the shipment of critical commodities such as grain. Most of central Europe was even more fragmented, and the burden of feudal tolls was such as to make traffic shun the rivers (though they were more navigable than those of France and Britain) and opt for the costly and uncomfortable alternative of land transport.

The internal free trade of Britain

This large market was also the richest in purchasing power per head. Insofar as one can estimate the incomes of mid-18th-century Europe, the average Briton would seem to have been receiving some £12 to £13 per year (perhaps $300 in today's money), half again as much as a Frenchman and more, in proportion, as one moves eastward or southward to the poorer nations of the Continent. The greater purchasing power of the Englishman was shown in his diet—he ate white wheaten bread and much meat, where the Frenchman ate dark bread and soup; and in his consumption of manufactures—he wore leather shoes rather than clogs, wool rather than linen, and used some 13 kilograms of iron per head (as against 2.7 kilograms in France) in the late 1780s. This was far more than even the rapidly growing domestic industry could supply, and the difference was made up by imports from Russia and Sweden.

The purchasing power of the British population was, in its mass, an important stimulus to agricultural and industrial production. The effect was reinforced by the relatively wide distribution of wealth, both socially and geographically. Unlike most European nations, Britain was not polarized between a small class of wealthy consumers of highly individualistic, labour-intensive luxury products and a large mass of poor peasants with little to spare for manufactures and those only the highly differentiated products of local industry. British society was more complex in its structure, with a large and heterogeneous middle class, and the peasantry was, in many parts, fully integrated into the national market. This, in turn, reflected the precocious commercialization of agriculture, which learned early to specialize in cash crops, above all for the great London market. Specialization means dependency on the outside, and one of the striking features of the English countryside in the 18th century was the extent to which not only itinerant peddlers but even fixed shops (which, of their nature, are evidence of a substantial and continuing trade) were catering to the rural population. All of this conduced to the breakdown of regional peculiarities and the homogenization of taste, which promoted a demand for the kind of standardized product that lent itself to machine production.

Social homogeneity

Coal and iron. This domestic demand, reinforced by a large and growing demand for British products abroad, placed heavy strains on the existing techniques and organization of production. One of the first industries to feel the pressure was mining, where the deeper one dug, the more difficult it became to drain the infiltrating underground water from the pits. Coal mines especially were in trouble. In one colliery, 500 horses were required to raise the water, bucket by bucket. The solution was to replace the animals with tireless machines. In 1698 Thomas Savery introduced his "fire engine," a steam pump rather than a true prime mover; and then, in 1705, Thomas Newcomen invented his atmospheric steam engine, which worked a piston up and down a cylinder, alternately raising and lowering another piston that did the actual pumping. By 1767 there were 57 of these in the Newcastle coal basin alone, supplying some 1,200 horsepower, and there were probably hundreds of them in operation throughout the kingdom. The Newcomen engine was simple and sturdy. Some of them remained in use for half a century or more. But it was wasteful of energy, and as thus largely confined to the coal pits, where fuel was almost a free good. James Watt's invention of the separate condenser (patent 1769; first commercial application, 1776), cut fuel consumption 75 percent, freed steam power from this constraint, and made possible the industrial city; and it was Watt again who first succeeded (patents of 1782 and 1784) in converting the reciprocating (back-and-forth) movement of the piston into rotary motion, so that steam could be used not only to pump mines or blow furnaces but also to drive the wheels of industry.

Another industry that felt the strain of increased demand was ironmaking. Here, the most acute bottleneck was the shortage of cheap fuel and carbon. Since England had long since cut down most of its timber, repeated efforts were made to use coke in smelting, but the necessity of mixing the fuel with the ore in the furnace made it difficult to cope with the impurities in the coke. It was not until 1709, when Abraham Darby, a Quaker ironmaster at Coalbrookdale, Shropshire, had the good fortune to work with a fairly clean mix that coke smelting became a commercial possibility. Even so, other ironworks found it hard to follow suit, and it was not until more powerful bellows became available in the 1760s that the technique spread widely. By the end of the century, the old charcoal smelting had almost entirely ceased, and British production of pig iron, once also stagnant, was growing rapidly: approximately 25,000 to 68,300 tons from 1720 to 1788 (1.5 percent per year) to about 244,000 in 1806 (7.3 percent per year).

Effect of the use of coke on ironmaking

Pig iron, the product of the blast furnace, has serious limitations for industrial use. It is brittle, cannot be worked, and must be cast. It is suitable for objects such as pots or pipes that need not take pounding or stress, but it cannot serve to build machines, frame buildings, or carry vehicles. For these purposes, before the age of cheap steel, only wrought iron would do—that is, iron refined to eliminate the carbon imparted by the smelting process. The Industrial Revolution in iron manufacture, therefore, was not possible until a way was found to use coal here too. The breakthrough came in 1783–84, when Henry Cort patented the combined technique of puddling followed by rolling. The first decarburized the pig; the second squeezed out the remaining dross and forced the hot metal into the desired shapes.

The putting-out system. The scale of enterprise in coal mining and iron manufacture was already substantial even before the Industrial Revolution. The dependence on bulky raw materials required these branches to concentrate in those areas dictated by nature, and the technology called for a heavy investment in fixed plant and equipment. By contrast, light industries, such as textiles, were widely dispersed and relied extensively on the distribution (putting out) of raw materials to cottage workers, who turned them into semifinished and finished products. The so-called putting-out system offered the merchant-manufacturers two important advantages. First, it shifted the burden of fixed costs to the workers. All the merchant needed was a room to stock materials and finished goods. When the market was bad, the putter out simply held off, with few overhead charges to pay. Second, it gave the merchant-manufacturer access to a large pool of cheap labour. Country dwellers would work for far less than urban, not only because the cost of living was lower but also because the rhythm of agricultural labour was uneven and left long periods of relative underemployment that could be profitably used to eke out the income from the land.

The putting-out system marked a major advance over the artisan workshops of medieval towns. By freeing enterprise from the constraints of town guilds, it opened the way to growth and technological innovation. Britain's precocity in this regard (by 1400, over half the wool cloth made was woven in the countryside) changed what today would be called an underdeveloped country, exporting its raw wool to manufacturing centres abroad, into Europe's biggest producer of cheap textiles for export.

Textiles and machine technology. The growth of demand in the 18th century pushed rural putting out to its limits, especially in spinning, where it took four or five persons to keep one weaver supplied with yarn. Given the high cost of land transport, the extension of the radius of labour recruitment increased disproportionately the cost of distribution of materials and collection of finished work. Nor was it possible to get more work from those already engaged by increasing wages; on the contrary, the worker tended to set his task by his need, and increased payment only meant that he could earn what he needed with less work. At the same time, the tempation to embezzle the raw materials increased, at the expense of the quality of the final product.

It was in this context of frustrated entrepreneurial opportunity that manufacturers sought for some way to bring the labour force under control. One way was to concentrate work under one roof, so that the employer could supervise performance. This solution, however, was not feasible in the context of the old technology. So long as enlarged workshops had to rely on human power and on the same machines used in rural cottages, they could not compete, for they were burdened with higher fixed costs and had to pay their workers more to lure them into a disciplined environment. The answer was found in the invention of machines that grew to be too large and costly for cottage work—so big, indeed, that they had to be driven by animal or inanimate power: the spinning jenny, by James Hargreaves in 1765; the water frame, by Richard Arkwright in 1769 (the prototype of all modern ring spinning devices); and the mule (so-called because it was a hybrid of the two), by Samuel Crompton in 1779.

Machines for spinning

This sequence of inventions shifted the bottleneck from spinning to weaving, and, for a short while, weavers lived in unaccustomed affluence trying to keep up with the output of cheap yarn. But in 1787 Edmund Cartwright invented a machine loom, which, though it took some two decades to perfect, inexorably displaced the handloom weaver. The power loom was applied first to the coarser fabrics, for only the tougher yarns could stand up to the irregular motions of the shuttle. The machine builders, however, learned to solve this problem; and by the 1820s handloom weaving was a dying craft, and its practitioners were among the most pitiful victims of technological obsolescence.

The innovations in textiles were applied first in cotton, primarily because cotton fibres lend themselves well to mechanical handling. Wool, an animal fibre, is less uniform and tractable. It also was more costly than cotton and constituted a much larger share of total costs; so that the potential cost reductions offered by mechanization were smaller. Even so, the innovations in spinning offered gains in labour productivity of the order of over 100 to 1; those in weaving, up to 10 to 1; so that mechanization came to wool also, lagging about two decades behind cotton, and eventually extended to all the textiles. The effect of prior mechanization in cotton is shown by the change in the relative position of the two industries. Between 1769 and 1829 the value of wool exports increased from £3,897,000 to £5,362,000, while that of cotton grew from £212,000 to £37,269,000.

Similar changes took place in other industries: paper

manufacture, metalworking and woodworking, chemicals, machine building itself. The principles of mechanization were susceptible to wide ramification, and the factory context tended to promote change by restoring the link between enterprise and actual production that had been broken by putting out. In chemicals, more than in other branches, the lessons of science found application in industry. Here, too, the initial motivation came from a bottleneck: the high cost of washing, dyeing, and finishing the flood of fabrics engendered by mechanized spinning and weaving. The great innovations here consisted of the substitution of abundant mineral for scarce vegetable or animal resources: the French chemist Nicolas Leblanc's substitution of coal and salt for seaweed and wood in the manufacture of soda; bleaching by chlorine compounds instead of milk and sunshine; washing of wool with soap rather than urine. The manufacture of these new industrial compounds gave rise to considerable waste; the effort to dispose of these noisome by-products without damage to litigious neighbours, plus the desire to recover some of the potentially valuable components of the waste, proved one of the most powerful incentives to innovation in the chemical industry. Seen together, these technical changes made possible substantial gains in industrial output, overall product, and per capita income (see Table 1).

Table 1: Indicators of Economic Growth in Great Britain (1801–51)

	national income (£000,000; current prices)	population (000)	per capita income (£)	income from manufacture, mining, and building (£000,000)	share of manufacture, mining, and building (percentage)
1801	232.0	10,686	21.7	54.3	23.4
1811	301.1	12,146	24.8	62.5	20.8
1821	291.0	14,206	20.5	93.0	32.0
1831	340.0	16,368	20.8	117.1	34.4
1841	452.3	18,551	24.4	155.5	34.4
1851	523.3	20,879	25.1	179.5	34.3

Source: Phyllis Deane and W.A. Cole, *British Economic Growth, 1688–1959* (1962); B.R. Mitchell and Phyllis Deane, *Abstract of British Historical Statistics* (1962).

Changes in transport. The growth of a modern industry in Britain generated a rapid increase in goods traffic. For all of Britain's natural advantages, the Industrial Revolution could not have taken place had there not been a concomitant improvement in facilities of transportation. Canals can be traced in Britain back to the 16th century, but the first truly industrial canal was the Sankey Canal (1757) from the Mersey to St. Helens, Lancashire, cut to bring Cheshire salt to Lancashire coal. This was followed in 1761 by the waterway that usually marks the beginning of the so-called canal age. This was the canal of Francis Egerton, the 3rd duke of Bridgewater, from his coal mines at Worsley to Manchester. By 1815 a network of some 2,200 miles crisscrossed the country, along with 2,000 miles of improved river navigation. The most prosperous links in the network were those that carried the heavy, bulk commodities of industry—coal, above all.

Canals and land transport

For many commodities, however, as well as for persons, only land transport was fast enough. The road network inherited from the pre-industrial economy was manifestly unable to handle the increase in traffic, and restrictions on vehicles (the imposition of wide wheels, for example, that would cut less into the road surface) were at best a palliative. Here, the answer was found, as in canals, in private enterprise. In the first half of the 18th century, Parliament averaged eight road acts a year, most of them authorizing new turnpikes; from 1750 to 1790, almost 40 acts a year were passed; and, from 1791 to 1810, the figure jumped to 55. Even so, the wear and tear played havoc with traditional road surfaces, and it was not until John Loudon McAdam and Thomas Telford introduced their new hard-packed surfaces in the 1790s that roads began to be built that could stand industrial traffic. By 1815 there may have been 1,000 miles of highway built to this standard.

All this time, the pressure built up to speed communication and travel. Coaches were built to make better times; horses were driven harder; connections were improved so that the trip from London to Birmingham, which took two days in the 1740s, was cut to 19 hours in the 1780s. By the early decades of the 19th century, these efforts had reached a limit, and the newspapers of the period are sprinkled with gruesome stories of horses snapping their legs or dying of exhaustion on the fast routes.

Railways

It was this apparently insatiable demand for speed that made the railway an instant success for passengers as well as goods. The use of rails to facilitate the movement of heavy vehicles went back centuries in mining districts; rails diminish considerably the force required to haul a given load. What made the modern railway, however, was the invention of locomotive traction introduced in 1825 on the Stockton and Darlington Railway line (primarily for coal haulage) and then, triumphantly, in 1829 on the new Liverpool and Manchester Railway. This was the beginning of the railway era. The new line had been intended for the transport of goods; but to everyone's astonishment passenger revenues exceeded freight receipts. By 1845, 1,000,000 persons were using the London and Birmingham Railway line alone; by 1850, 6,559 miles of track were in service.

The railway was probably the most impressive and significant technical innovation of its time. Not only did it enable the economy to move the unprecedented volume of goods required and produced by the new techniques, but also its demand for iron (for rails, engines, rolling stock), coal, wood, brick, and other materials stimulated other industries. The requirements of a railway roadbed, in particular the need for level track, inspired some of the most spectacular engineering achievements of the age. The construction of the roadbed (including tunnels, bridges, viaducts, etc.) called for large gangs of labourers and proved to be a laboratory for new techniques of industrial management. Similarly, the unprecedented costs of railway construction called for new methods of mobilizing capital and pushed Britain in the direction of corporate business enterprise with limited liability.

All of these connections of the railway with the rest of the economy made it the most important single determinant of the level of business activity. At the peak of the railway boom of the 1840s, the industry employed some 300,000 men in construction alone—more than in the entire cotton industry—and absorbed about half the country's total investment; that is, between 5 and 7 percent of the national income. It was also a potent force for geographical mobility and contributed thereby to the efficiency of the labour market and the democratization of society—this, in spite of segregation of passengers into different price classes.

Improvements in agriculture. The revolution in industrial technology was accompanied by important changes in the agricultural sector. Indeed, these were indispensable to the expansion of manufacturing; the new industrial work force was cut off from the land and depended for existence on the production of a food surplus and on such victuals as could be purchased abroad. The problem was aggravated by a rapid growth of population, beginning in the middle of the 18th century, and by the constriction of overseas supplies during the period of the French Revolutionary and Napoleonic Wars (1792–1815).

At the start of the 18th century, Britain, in years of good harvest, was still able to send food abroad. Agriculture was highly commercialized and specialized, and the techniques of cultivation were among the most advanced in Europe. Some scholars have even argued that the innovations of the previous period—water meadows plus mixed farming, new crop rotations, marling—were more important than the better-publicized advances of the so-called agricultural revolution of the 18th century.

New farming techniques of the 18th century

Over the course of the 18th century, the techniques of agriculture continued to advance. Certain names stand out: Jethro Tull, who advocated drilling seed rather than sowing it broadcast and who also advocated light iron horse-drawn implements instead of cumbersome wooden tools drawn by slow oxen; Charles ("Turnip") Townshend, 2nd viscount Townshend, who learned about turnips and

clover in Holland and helped popularize them in England; Robert Bakewell (1725–95), the most prominent livestock breeder of his day, the creator of four new breeds of sheep selected for meat instead of wool; Thomas William Coke of Holkham, the personification of the gentleman farmer and chief exponent of the so-called Norfolk system of rotation. But what really made an agricultural revolution possible was not these few public figures but the large number of anonymous improvers, adopting and adapting the new methods with an alacrity and vigour that reflected both the enhanced market incentives and a veritable agronomania, which found expression in and was nourished by a flowering of agricultural societies and publications. In all of this, the emphasis was on rational accounting procedures that could provide the basis for an exact comparison between old ways and new: the reports of a contemporary observer such as Arthur Young consisted in large part of records of outlays and receipts, county by county, farm by farm.

The introduction of these new techniques encountered a serious obstacle in much of England in the prevailing pattern of dispersed, fragmented land tenure and open fields subject to common rights of usufruct. These arrangements went back to the Middle Ages or even beyond. They reflected, no doubt, an effort to equalize the quality of soils allocated to the different members of the village community: each received a piece of each of the different kinds of land cleared for cultivation. The common rights of usufruct followed: this was the only feasible way to permit each man to pasture his livestock on the stubble after harvest. Common rights of pasture, in turn, implied concurrent sowing and harvesting, hence a common pattern of crop rotation, which usually entailed one year of three in fallow. This wasteful system locked the farmer into a vicious circle of low yields—insufficient food for wintering livestock—low output of dung—inadequate fertilizer—low yields. By contrast, introduction of clover or root crops not only did away with fallow but provided additional food to keep the animals over the winter, to build up the herds and flocks, and to accumulate stocks of dung for the growing season.

The enclosure movement

The attempt to free land from these communal servitudes and regroup holdings gave rise to what was known as the enclosure movement. In village after village, in an area covering perhaps one-third of the arable land in England, the major landholders agreed to regroup holdings into contiguous parcels that could then be "enclosed" by hedges or ditches and farmed as the owner saw fit; at the same time, meadowland and wastes held in common were divided among the landholders in proportion to their holdings. This was not the first such wave of enclosure in English history: perhaps half of the arable land had already been converted by 1750. These earlier enclosures—in the 16th century especially—had aimed at converting arable land into large sheepwalks, to meet the growing demand for wool. The enclosures of the 18th century, however, were intended to increase cultivation.

Although enclosures were perhaps the greatest single factor in permitting and encouraging the diffusion of the new agricultural techniques, they entailed a heavy charge on the smaller landholders, who could not easily afford the legal costs and outlays for surveying, hedging, ditching, and the like. Even worse was the situation of the landless cottar or bordar, who lived by a kind of tolerance on common land, was permitted to pasture a few animals, and worked as a day labourer in the fields of his propertied neighbours. Enclosure settlements were scrupulous in their protection of the rights of those who owned land but were equally rigorous in their exclusion of those who did not. As a result, many small holders were forced to sell their parcels to those who could afford the costs of enclosure; and many more landless labourers had to leave their homes and take service elsewhere. For a long time, it was thought that these uprooted victims of enclosure provided the basis of the factory labour force. This has been disproved by the census data, which show the population of areas of enclosure growing throughout this period. This is exactly what one would expect of an institutional change that increased the demand for farm labour. Even so, agricultural employment could not keep up with increasing manpower, and it was the surplus population of the countryside that flowed to the cities. Enclosure selected many of those who chose to go; but, on balance, the agricultural revolution it helped make possible slowed the course of rural emigration.

There are no official figures on British farm output in this period, but it is estimated that wheat production, for example, almost doubled from 1700 to 1820—13 million to 25 million quarters. Even so, the agricultural sector could not supply the growing food needs of the country, and, from the 1770s on, Britain was on balance an importer of grain. To be sure, imports represented a small fraction of overall wheat consumption—less than 5 percent of home demand in the period from the end of the French Revolutionary Wars to the abolition of the Corn Laws (tariffs on imported grain) in 1846. Yet, the proportion would have been significantly larger had it not been for these protective duties, which constituted a subsidy to British farming and kept British prices higher than they would have been in a free market.

Abolition of the Corn Laws

The abolition of the Corn Laws and the introduction of almost complete free trade in manufactures in 1845–49 marked the substantial completion of Britain's Industrial Revolution and its economic pre-eminence at midcentury. The first signalled the victory of the manufacturing interest over the agricultural: as early as the second decade of the century, employment in industry exceeded that in agriculture; by 1851 it was twice as large, as was the share of manufacturing in the national product. If one added trade, transport, and other predominantly urban activities to the balance sheet, the disparity would be much greater. Understandably, the modern sector refused to go on paying tribute to the land.

The second was testimony to Britain's confidence and superiority in the world market; no one could undersell its wares (some 93 percent of British exports were manufactures). At this point, this small country, with less than 10 percent of the population of Europe, accounted for two-thirds of the coal mined in the world, half the make of iron, half the market of cotton cloth, and over one-quarter of the goods in international trade. It was a moment of dominance without parallel before or since.

Continental Europe. *The Continent until 1815.* The other countries of Europe actually took a long time effecting their industrial revolutions. The rapidity with which some of them moved once the new technology took hold—Germany is the usual example—is deceptive. The hard job was to start moving. By comparison with Britain, the Continent was severely handicapped: geography, political boundaries, and a legacy of medieval tolls were barriers to the movement of persons and goods; forests were relatively abundant but coal scarce; incomes were lower; labour was cheaper and labour-saving devices were thus less advantageous; society was more rigidly divided by status and occupation; it was more rural, and country dwellers made poor or tight-fisted consumers; social attitudes were less favourable to business enterprise. In the lands east of the Elbe, the persistence of serfdom locked the greater part of the population into agriculture, so that the potential supply of labour to industry was drastically curtailed; moreover, since unfree men do not work so well as free, productivity was low, and the people were impoverished. Even in the freer lands to the west, however, the rural economy could not be compared with that of England. Subsistence farming predominated; there was considerably less regional specialization; techniques were less productive, and there was nothing like England's fascination with improvement. Commerce and industry had scarcely begun to penetrate the countryside.

Government sponsorship of manufacture

By the mid-18th century, even earlier in some countries, the governments of Europe were seeking systematically to promote industrial development. The reasons were primarily political: industrial growth and economic prosperity generated revenue and employed men, and money and men were the sinews of war, hence of power. The measures varied with and are a good reflection of the needs and shortcomings of these economies: subsidies in the form of money, gifts, low-cost loans, exemptions from taxes, use

of public buildings, or free machines; technical assistance and the imposition of compulsory standards and quality control (particularly important in France); monopoly privileges to innovators of new products or techniques; an assured supply of labour, by assignment of serfs, prisoners, vagabonds, or such public charges as orphans and the residents of poorhouses (especially in central and western Europe); exemption of the work force from military service and the plant itself from the burdens of quartering troops; special privileges to immigrant artisans and entrepreneurs (Huguenots in Berlin, Germans in Russia, Jews throughout central Europe); an assured market via government or army orders or compulsory purchase by such unfree subjects as Jews (Germany especially); the award of places and honours to industrialists and of patents of distinction to their products. In some fields, particularly those linked to armament (powder works, arsenals) and to the supply of luxuries to the court (glass and mirror manufacture, tapestries and carpets, porcelain), the state itself became entrepreneur. In other branches, the state relied on private initiative and, when this failed to respond to the incentives offered, might order aristocrats or court purveyors or Jews or such others as were in no position to refuse to start putting their money and efforts into manufacture.

Much of this development from above was misdirected. Technical knowledge and managerial talent were in short supply, and the effort to force progress inevitably led to misallocation of resources and the manufacture of goods of poor quality at high cost. Subsidies only encouraged waste and inefficiency, while monopoly privileges eliminated the competition that might have compelled better performance. The difficulty was compounded by the preference of most rulers and their officials for large, centralized establishments—visible symbols of industrial achievement. The "manufactories," so called to distinguish them from true factories, were not centrally powered and were ordinarily equipped with the same devices used in conventional workshops. They had, therefore, no technical advantage over dispersed cottage manufacture yet were burdened with higher capital costs and were far less attractive to workers, who resented supervision and discipline—hence the need to assign labour, which was invariably sullen and unproductive. Most of these hothouse enterprises collapsed as soon as the state withdrew its support, and the setback was proportional to the amount of support. In Prussia, which was the greatest exponent of forced development, the death of Frederick II the Great, in 1786, was followed by a hecatomb of bankruptcies and liquidations.

This is not to say that the whole effort of state sponsorship of industry was pointless. The privileged manufactories left behind a legacy of knowledge and skills. Many of the men who were trained in them constituted a second generation of entrepreneurs, less spoiled, more hardheaded, and far more successful.

On the Continent, as in Britain, the real progenitor of modern industry was the putting-out system. It had always been inhibited by the privileges of urban manufacturing centres and the efforts of the state to control the standards of production; but, beginning in the late 17th century, industry gained a foothold in the French, Walloon, Rhenish, and Swiss countrysides. In France, the prosperity and influence of this partially illicit industry was recognized by a law of 1762, which legalized the fait accompli. By this time, the efforts of merchant-manufacturers to reach farther afield for their labour (Norman clothiers were drawing some of their yarn from as far away as Picardy) were producing the same kinds of bottlenecks that had moved British manufacturers to turn to machinery and the factory, and there is no doubt that the same changes would have occurred on the Continent had the economy simply continued to develop along these lines. But this process was interrupted, first, by the British innovations and, then, by the French wars of 1792–1815. The first changed completely the conditions of international competition. It was at once apparent that any nation that wanted to maintain itself as a power, political as well as economic, would have to learn the new technology. The second, however, worked against this by cutting much of the Continent off from Britain for a generation and diverting resources into familiar but profitable techniques rather than more hazardous innovations.

The spread of British techniques

The extension of the Industrial Revolution to the Continent may be viewed as a process of cultural diffusion. The continental countries already had a pool of skilled artisans who might learn and apply the new techniques—clockmakers, instrument makers, millwrights, joiners, metalworkers. Even so, they needed considerable help to bring them to the British level. This was accomplished by missions of inquiry to Britain or by placing workers in British plants, by a certain amount of industrial espionage, by the purchase of British machines (usually illegally), and by engaging Britons to work in continental enterprises.

The first fruits of this diffusion were already seen before the French Revolution: a few steam engines, including some built in Europe; some thousands of mechanized spindles for spinning cotton; the first, abortive attempt at smelting iron ore with coal at Le Creusot in France (1785). During the French wars, the interruption of British imports permitted certain branches, textiles especially, to develop in a kind of splendid isolation—that is, to modernize with old-fashioned equipment. Such crucial branches as engineering and iron manufacture, however, made little headway. The one area in which isolation paid was chemicals, where the French responded to the shortage of alkalis by developing the Leblanc soda process.

The continental revolution, 1815–50. The return of peace brought both opportunity and peril: on the one hand, communications with Britain were open once again, and, on the other, British manufactures were so much cheaper and often better than continental wares that the infant industries of the Napoleonic era were menaced with liquidation. Most of the continental countries responded by setting up prohibitive tariff barriers; at the same time, both government and private enterprise bent every effort to import British skills and experience or to spy out the secrets of British technology. To combat this potential loss of superiority, the British had long since passed laws against the migration of skilled artisans or the export of machinery; but these could do no more than impede the process. By 1825 Britain's attempt to maintain a monopoly of the new technology was clearly a failure, and the ban on the emigration of artisans was lifted. Machines took longer, but in 1842 pressure from a machine-building industry whose capacity had outstripped home demand persuaded the Parliament to annul the prohibition on exports.

Determinants of continental changes

The rate of assimilation and adoption of the new technology in continental Europe varied widely. Among the more important determinants were (1) the degree of economic and technological backwardness; (2) the degree of exposure to British competition (too much could be fatal, but a certain amount could be stimulating, as in Belgium and Switzerland); (3) relative factor costs (if labour was cheap enough or capital too scarce and dear, it did not pay to install machines); (4) the political pressures for modernization (the rulers of smaller states, such as Belgium, Switzerland, and Sweden, not caught up in the competition for power, were willing to let economic development take its course); (5) the political pressures for conservatism (old-fashioned landed magnates in such countries as Austria and Russia correctly perceived industrialization as socially and politically subversive); (6) the resource base (other things equal, a goodly supply of coal, as in Belgium and Westphalia, was a potent incentive to the adoption of the new technology); (7) the size of the market (the unification of most of Germany into a customs union [Zollverein] by 1833 gave a powerful impetus to trade and regional specialization of industry); (8) the character of demand (peasants made poor customers for manufactures); and (9) the ability of the society to generate entrepreneurs (weakest in eastern Europe but also constrained in countries such as France by hostility to unbridled competition and by the conservatism of family enterprise).

The years from 1815 to 1850 saw several of the continental countries initiate their own industrial revolutions. Belgium was first, in the 1820s and '30s, with a marked emphasis on coal mining and metallurgy. France followed and, being poor in coal, concentrated on textiles and other

light industries. From the 1830s on, the Swiss moved along the same path. Finally, Prussia, with large coal resources in Silesia and Westphalia, developed from the 1840s on a pattern similar to Belgium's. In all of these, railway construction was an important stimulus to development. Belgium, compact and densely populated, built its trunk network by the early 1840s. Germany was almost as quick, so that by midcentury one could cross the entire country by rail, east–west or north–south. France, by contrast, had only stubs of lines radiating from Paris and isolated stretches in the provinces. In the rest of Europe, the new technology was still exceptional, stunted by material and institutional obstacles.

The middle decades, the 1850s and 1860s. The decades of the 1850s and '60s were a period of catching up for continental Europe—not that Britain was standing still. Easy money favoured expansion: the effect of the gold discoveries in California and Australia (1848–49) was multiplied by the readiness of credit institutions to lend in a highly liquid market. The contrast was especially marked with the crisis years of the 1840s, when bad harvests, the collapse of a speculative boom in railway shares, and, finally, an epidemic of political revolutions in 1848 had first contracted the economy and then just about brought it to a standstill. Now confidence reigned everywhere, and promoters hastened to bring new commercial and industrial enterprises to the attention of investors. A striking feature of this renewed expansion was the role played by joint-stock, limited-liability corporations. These were not new; they went back to the chartered trading and colonial companies of the 17th century. But they had always been exceptional, subversive of a commercial morality that held a businessman personally liable for his losses "to the last shilling and acre." They had required the express authorization of the state and had been reserved for those few activities that could not be handled even by the collective resources of a partnership. Railway construction was one of these, and the railway did more than anything else to naturalize the limited-liability corporation. The principle was susceptible, however, of much wider application, the more so as technological advance and economies of scale increased the threshold investment required to found competitive enterprises. The contribution of limited liability was far more important on the Continent than in Britain, which had built its industrial establishment from the ground up, plowing back profits into firms that began small and grew with the changing technology. Britain was also richer and possessed the most efficient money and capital markets in the world, so that industrial enterprises could count on generous bank credit, nominally short-term, but, in fact, renewed to cover medium- and long-term investments. By contrast, the follower countries across the Channel, which had to bridge the technological gap, were confronted with the task of paying for costly installations, not only with the industrial profits of yesterday but also with such resources as could be mustered from outside the industrial sector. In such circumstances, limited liability and joint stock were indispensable, and institutions had to be developed to take up the task of mobilizing capital and channelling it to potentially profitable ventures—hence, the importance of the corporate investment bank, already successful in Belgium but taking a new departure in the 1850s with the establishment of the Crédit Mobilier in Paris (1852) and the Bank für Industrie und Handel in Darmstadt (the Darmstädter Bank, 1853). These were followed by (and, in numerous instances, helped found) similar development banks throughout Europe. Some of them concentrated more on speculative promotions than on industrial development; the Crédit Mobilier itself, eponymous model for the new species, foundered in 1867 because of its vain effort to keep up profits and appearances by stock-market coups. Others worked better, especially in Germany, where the rapid growth of capital-intensive heavy industry, above all in metallurgy and mining, provided a large field of opportunity.

The role of limited-liability corporations

Free trade

The rapid economic expansion of these years set the stage for a major departure from the conventional wisdom of high protection. The key move was not Britain's unilateral step toward free trade in 1845–49 but rather the Anglo-French Treaty of 1860, which was swiftly followed by a series of similar agreements between the leading European economies, including the Zollverein. These were accompanied by reciprocal legislation to facilitate capital movements across national borders and by international monetary agreements providing for standardization and simplified exchange of currencies. This was not yet complete free trade, but it did bring closer the kind of international market that would, in turn, promote technological rationalization and diffusion. The Cassandras who predicted that the protected industries of France and the Zollverein would simply be crushed under the weight of British imports were proved wrong; the most efficient and strongest continental enterprises invested heavily in new plants and did better than ever. There was, however, a purge of smaller, marginally inefficient firms unable or unwilling to meet the new competition.

Improved transport

This purge of the old technology was further hastened by the vast improvement in transport technology that marked these years, in particular, by the competition of the trunk rail network of western and central Europe. The track mileage rose sharply, but more important than this was the completion of key connections and the consequent growth of passenger and goods traffic: 65,000,000 to 169,000,000 tons in Britain from 1855 to 1870; 4,000,000 to 44,000,000 in France from 1850 to 1870. Nothing did more to break down the autonomy of regional markets and promote the mobility of persons; internal migration, up to then confined largely to short intraregional movements, began now to develop along interregional routes. The railway, in combination with the steamship, also moved a rising current of transoceanic migrants—108,000 per year from Europe to the United States, 1830–49; 218,000 per year, 1850–69. This exodus, as much as modern market competition, hastened the demise of old trades in low-wage rural areas.

Table 2: Growth of Railway Networks in the 19th Century, Selected Countries
(in kilometres)

	1840	1850	1860	1870	1880	1890	1900
Great Britain	2,410*	9,791	14,595	21,826	25,046	27,811	30,063
France	499†	2,915	9,167	15,544	23,089	33,280	38,109
Belgium	325	625	747	869	2,724	3,249	4,060
Germany	549	6,044	11,633	19,575	33,838	42,869	51,391
Russia‡	26	601	1,589	11,243	23,857	30,957	48,107
United States	4,534	14,515	49,292	85,139	150,717	268,409	311,094

*United Kingdom. †1841. ‡European Russia only.
Source: B.R. Mitchell and Phyllis Deane, *Abstract of British Historical Statistics* (1962); France and Belgium, Statistical Annuals; W. Woytinsky, *Die Welt in Zahlen*, V. (1927).

Although the middle decades of the 19th century were dominated by the extension of known techniques to new enterprises and areas and within existing enterprises, technical advance did not cease. For one thing, considerable improvement was built into the very process of expansion; it is hard to build a new edition of an old model without incorporating improvements. For another, no technique or machine is usable without some adjustment to the specific conditions in which it is to be used, and this adjustment often entails improvements. In the meantime, the pressures of demand continued to stimulate invention. An excellent example is steelmaking, where the traditional cementation and crucible techniques could produce only small pieces at high cost, used in shears, razors, swords, watch springs, and the like. The development of high-speed machinery and vehicles, however, created a need for steel in far greater sizes and amounts than customary. From the 1830s on, then, metallurgists in several countries bent their efforts to the problem. The definitive solution was found at about the same time by William Kelly of the United States (*c.* 1851) and Henry Bessemer of England (*c.* 1854), who was looking for a way to make a better cannon: by blowing air through a furnace containing molten pig iron, the oxidation of the carbon would proceed without the application of any external heat; hence, Bessemer's claim to make steel "without fuel." Though taken up eagerly, the new technique needed some additional touches before it was

Cheap steel

susceptible of general application: first, Robert Mushet's clever trick of burning off all the carbon and then adding it as desired made possible a standardized product, and, much later, the Thomas–Gilchrist process (1878–79) enabled steelmakers to work with phosphoric iron ores and opened the greater part of the world's iron resources to the new technique. In the next decade, a second way was found to produce cheap steel, this time with even closer quality control: first, a new "regenerative" oven invented by William Siemens (mid-1850s) yielded much higher temperatures, so that iron could be refined in huge open hearths; and here again, it was a second inventor, Pierre Émile Martin of France, who worked out the technique of adjusting the mix by the admixture of scrap so as to yield a standard product (1864). The whole story of cheap steel is almost an ideal case study in the dynamics of invention: (1) the importance of serendipity, (2) the role of multiple invention, and (3) the incompleteness of initial solutions and the indispensability of subsequent improvement.

The chemical industry

The chemical industry also moved ahead; in the 1850s and 1860s artificial, synthesized dyes were developed, first in England, then in Germany. Much more than in steel-making, however, technical advance in chemistry depended on scientific knowledge and systematic experiment: artificial dyes were only the first application of the new science of organic chemistry. These first dyes, then, were a promise not only of what would rapidly become one of the most creative and flourishing branches of industry but also of a new approach to invention.

By 1870 the more advanced industrial countries of continental Europe had reached a point at which they could compete with Britain on reasonably equal terms. The scale of their enterprises was still smaller, as was the place of industry in the larger economy. Yet, they had assimilated the essential techniques of machine manufacture and adapted them to their circumstances; and, where before they had been almost exclusively learners, they were now generating innovations of their own.

Social consequences of the Industrial Revolution. The question of the social consequences of the Industrial Revolution has been hotly debated since its effects were first noted in the early 19th century, and the issue is still alive. The split has been political, with the right usually giving a favourable judgment, the left a negative one. The question is beset by difficult problems of interpretation, value judgments, and criteria, but the following general observations may be made concerning the impact of the Industrial Revolution in Britain.

Standards of living

First, wage data indicate that, on balance, those workers who shifted out of agriculture or pre-mechanized industry into the modern branches of industry improved their earnings significantly. The gap was largest for those who came from the most backward rural areas in western and southern Ireland or the Highlands of Scotland. Those who stayed behind in agriculture benefitted to the extent that labour could easily take advantage of job opportunities in manufacturing: rural wages were highest in those parts of England adjacent to the great industrial centres.

Second, the worst losers were workers in bypassed trades, such as handloom weaving. By 1830, tens of thousands of these were living on the edge of survival.

Third, real earnings fluctuated with prices, particularly food prices, and the level of employment, as they always had. Real wages were thus at their lowest during the period of the French wars, when wheat prices reached record heights, rose over the next 15 years, fell again in the depression of the late 1830s (despite falling prices), then moved back up from the trough of 1842. There is no agreement on what the composite trend was from, say, 1790 to 1850, though no one would argue for a marked shift either up or down; the consensus, however, is that real wages rose substantially from the middle of the century on.

Fourth, consumption of the major food staples seems to have increased, partly owing to new sources of nourishment (potatoes and fish, in particular). Consumption of textiles grew even faster, and the generalization of the use of washable undergarments thanks to cheap cotton fabrics constituted a major advance in comfort and health.

Fifth, on the other hand, housing deteriorated as thousands of migrants poured into swollen urban centres that could not grow fast enough to accommodate the influx. The population of 103 cities of over 20,000 (including London) rose from 3.5 million in 1821 to 6.8 in 1851 to 9.5 in 1871. (This, it should be noted, is a general characteristic of all industrial revolutions: because construction technology does not advance as rapidly as that of other branches, investment in housing is less profitable than other investments or can be made profitable only at the sacrifice of quality and space—this, both for systems of private enterprise and state planning.) The poor either jammed into old structures, filling every vacant space by building sheds in courtyards or converting cellars into dwellings, or moved into little jerry-built row houses that were slums before they were finished. Water supply and sewage disposal were grossly inadequate: the life of the urban poor was a compound of damp, filth, and nasty smells. All of these inconveniences were an old story; rural housing was often even more primitive; but 100 hovels scattered over the hillsides were a much less serious problem than back-to-back tenements. Density bred discomfort, disease, and neurosis. Urban death rates actually rose in Britain in the 1820s and 1830s.

Working conditions

Sixth, the conditions of work in the modern industrial sector varied widely. Generally, the larger and more modern the enterprise, the roomier, clearer, and safer the workrooms, the shorter the hours, and the higher the wages. The weaker, less-efficient firms had to save on their wage bill in order to compete. Over time, therefore, the construction of new plants and increasing scales of production yielded as a by-product a substantial improvement in working conditions. The worst feature of the new mode of production was the extensive reliance on women and children. In itself, this was not new: women and children had always worked hard and long in cottage industry. Nevertheless, the discipline of the factory (a discipline often imposed by strangers), the remorseless monotony of many of the tasks, and the physical hazard and discomfort of some of the new processes took a heavy toll.

By the first decade of the 19th century, the abuses of factory labour were already a cause of public concern in Britain. The government was at first reluctant to intervene, for fear of breaching the sanctity of private contracts, but, beginning with parish apprentices in 1802, then continuing with other children in 1819 and 1833 and with women (in mines) in 1842 and (in textile factories) in 1847, Parliament gradually increased its purview. The most important step was the passage of the Factory Act of 1833, which set a minimum age of nine years, limited the hours of children to eight per day, required employers to provide schooling for their young workers, and, most decisive, established the principle of official inspection to ensure compliance. Unfortunately, the law aimed at first to protect only workers in centrally powered textile factories, so that older modes of production were enabled to persist in abusive working conditions until the passage of more comprehensive legislation later in the century.

In general, the social consequences of the Industrial Revolution were not so uniformly bad as they have been made out but were far worse than they had to be. They reflected the unpreparedness of British society, the ideological commitment to laissez-faire, and the natural indifference of most beneficiaries of industrialization to the sufferings of those who paid the bill.

Across the Channel, observers of the British scene were inclined to congratulate themselves on being spared the "white slavery" of the cotton proletariat. Yet, their poverty problem was even more severe—though more widely distributed, hence, less visible. Their industrial labour force was paid less and worked harder, and more of their children died younger. Yet data on the consumption of necessities, such as wheat, potatoes, and textiles, show substantial increases per head over time; and, since these gains could not have been confined to the well-to-do, one must conclude that people were eating more and dressing better. Housing, as always, was a dark spot, with death rates varying widely between rich and poor neighbourhoods. Much depended on the level of economic activity:

the late 1830s were hard years, and the mid-1840s, with their bad harvests and business contraction, produced in some areas (Flanders, Silesia) a Malthusian peak in mortality. Things picked up thereafter.

Another test of the effects of industrialization is the hypothetical question: what would have been the condition of the population had there been no Industrial Revolution? It is nearly impossible to conjecture an answer. One approach is to consider the fate of those parts of Europe, or even of the United Kingdom, that did not keep pace with the new technology. These proved invariably to be the poorest areas—sloughs of overpopulation, endemic underemployment, and undernourishment. Like Silesia in 1844 or Flanders and Ireland in 1846, they had no reserves in time of want and fell easy prey to famine and disease. There was an autonomy to the population growth that began in the mid-18th century that made some kind of revolution in productivity indispensable if Europe was not to lose the slow gains of a millennium and sink back to the poverty of the Middle Ages. (D.S.La.)

ROMANTICISM AND REALISM

The legacy of the French Revolution. To make the story of 19th-century culture start in the year of the French Revolution is at once convenient and accurate, even though nothing in history "starts" at a precise moment. For although the Revolution itself had its beginnings in ideas and conditions preceding that date, it is clear that the events of 1789 brought together and crystallized a multitude of hopes, fears, and desires into something visible, potent, and irreversible. To say that in 1789 reform becomes revolt is to record a positive change, a genuine starting point. One who lived through the change, Liancourt, was even sharper in his vision when (as the story goes) he answered Louis XVI, who had asked whether the tumult outside was a revolt: "No, sire, it is a revolution." In cultural history as in political, significance is properly said to reside in events; that is, in the acts of certain men or the appearance of certain works that not only embody the feelings of the hour but also prevent other acts or works from having importance or effect. To take an example from English poetry, after Wordsworth and Coleridge's *Lyrical Ballads,* which appeared in 1798 and remained a long while obscure, the writing of heroic couplets in the 18th-century manner could be but a pastime—not poetry, as it had once been. A good many such "punctuation marks" can be noted in the text of cultural activity and in its context of political and social life.

To say, then, that the cultural history of the later modern age—1789 to the present—begins with the French Revolution is to discuss that Revolution's ideas rather than the details of its onward march during its first 10 years. These ideas are: the recognition of individual rights, the sovereignty of the people, and the universal applicability of this pair of propositions. In politics the powerful combination of all three brings about a permanent state of affairs: "the revolution" as defined here has not yet stopped. It continues to move the minds of men, in the West and beyond. The revolution is "dynamic" because it does not simply change rulers or codes of law but it also arouses a demand and a hope in every man and every people. When the daily paper tells of another new nation born by breaking away, violently or not, from some other group, the Revolutionary doctrine of the sovereignty of the people may be observed still at work after two centuries.

The ideas of the French Revolution

Cultural nationalism. The counterpart of this political idea in the 19th century is cultural nationalism. The phrase denotes the belief that each nation in Europe had from its earliest formation developed a culture of its own, with features as unique as its language, even though its language and culture might have near relatives over the frontier. Europe was thus seen as a bouquet of diverse flowers harmoniously bunched, rather than as a uniform civilization stretching from Paris to St. Petersburg, from London to Rome, and from Berlin to Lisbon—wherever "polite society" could be found, a society acknowledging the same artistic ideals and speaking French. In still other words, the Revolutionary idea of the people as the source of power ended the idea of a cosmopolitan Europe.

The "uniform" conception presupposed a class or elite transcending boundaries; the "diverse" implied a nation of citizens attached to their native soil and having an inborn and exclusive understanding of all that had been produced on it. In each nation it is the people as a whole, not just the educated class, that is deemed the creator and repository of culture; and that culture is not a conscious product fashioned by the court artists of the moment: it is the slow growth of centuries. Thus can be seen one of the motives of the post-Revolutionary passion for historical studies—the change to cultural populism as the replacement of a single horizontal, Europe-wide, and "sophisticated" culture by a series of vertical, national cultures, popular not only in their generality but also in their simplicity.

Beginnings of cultural nationalism

This new outlook, though propagated by the Revolution, began as one of those subdued feelings mentioned earlier, as undercurrents beneath Enlightenment doctrine. In England and Germany especially, a taste developed for folk literature—the border ballads, the legends and love songs of the people, their dialects and superstitions. Educated gentlemen collected and published these materials; poets and storytellers imitated them. Horace Walpole in *The Castle of Otranto,* Macpherson in *Ossian,* Chatterton in his forgeries of early verse, and Goethe in his lyrics exploited this new vein of picturesque sentiment. A scholar such as Herder or a poet-dramatist such as Schiller drew lessons of moral, psychological, and philosophical import from the wisdom found in the subculture of *das Volk.* The folk or people was not as yet very clearly defined, but the Revolution would shortly take care of this omission.

In France, where the Revolution occurred, the situation was somewhat different. There were no collectors of border ballads or exploiters of Gothic superstitions. France by 1789 had been for more than a century the cultural dictator of Europe, and it is clear that in England and Germany the search for native sources of art was stimulated by the desire to break the tyranny of the French language and literature. The rediscovery of Shakespeare, for example, was in part a move in the liberation from French classical tragedy.

Simplicity and truth. But it was also the expression of a genuine desire for simplicity and truth and for the release of feelings that the confidence of the Enlightenment in the power of reason had tended to suppress. Two 18th-century figures tapped this fount of emotion, Richardson and Rousseau. The novels of Richardson, in which innocent girls are portrayed as withstanding the artful seductions of titled gentlemen, might be said to foreshadow in symbolic form the struggle between high cosmopolitan culture and the new popular simplicity. These novels were best sellers in France, and Rousseau's *Nouvelle Héloïse* followed in their wake, as did the bourgeois dramas of Diderot, Beaumarchais's satirical comedies about the plebeian Figaro, and the peasant narratives of Restif de la Bretonne, to mention only the most striking exemplars of the new simplicity.

At the very centre of sophistication the simple life became a fad, the French court (including Marie-Antoinette) dressing up and playing at the rustic existence of milkmaids and shepherds. However silly the symptoms, the underlying passion was real. It was the periodic urge of complex civilizations to strip off the social mask and recover the happiness imagined as still dwelling among the humble. What was held up to admiration was honesty and sincerity, the strong and pure feelings of people unspoiled by city and court life. Literature therefore came to express an acute sensitivity to scenes of undeserved misfortune, of heroic self-sacrifice, of virtue unexpectedly rewarded—a sensitivity marked by tearfulness, actual or "literary."

Back-to-nature movement

This surge of self-consciousness about sophisticated culture has often been confused with an idealization of primitive man and attributed to Rousseau. But contrary to common opinion, the so-called back-to-nature movement does not at all echo the noble-savage doctrine of the 17th century. Rousseau's attack on "civilization," which evoked such a powerful response in the latent feelings of his contemporaries, goes with a characterization of the savage as stupid, coarse, and amoral. In Rousseau and his abettors, what is preached is the simple life. What nature

and the natural really are remains to be found by trial and error—the fit methods and forms of religion, marriage, child-rearing, hygiene, and daily work.

Populism. It is easy to see in these beliefs and sentiments (which often passed into sentimentality) additional materials for the populism that the Revolution fostered. Revolution, to begin with, is also an urge to simplify. The Revolutionary style was necessarily populist—Marat's newspaper was called "The Friend of the People." The visible signs that a revolution had occurred included the wearing of natural hair instead of wigs and of common trousers instead of silk breeches, as well as the use of the title of *citoyen* instead of *Monsieur* or any other term of rank. Now, equality coupled with sincerity and simplicity logically leads to fraternity, just as honest feeling coupled with devotion to the people leads to puritanism: a good and true citizen behaves like a moral man. He is, under the Revolutionary principles, a responsible unit in the nation, a conscious particle of the will of the sovereign people, and as such his most compelling obligation is love of country—patriotism.

With this last word the circle of ideas making up the cultural ambient of the French Revolution might seem to be complete. But in the effort to trace back and interweave the strands of feeling and opinion that make up populism, one must not overlook the first political axiom of revolutionary thought, which is the recognition of individual rights, often said to be natural rights as well. Their essence is a subject for political theory. Here their cultural role is of interest, and it can be stated in a very obvious way: individual rights generate individualism. The "ism" rightly suggests an attitude or doctrine—sometimes a passionate faith—asserting that every human being is an object of interest in himself, an end in himself. What is more, the truly valuable part of each individual is his uniqueness, which is entitled to thrive and develop free of oppression. That is why the state guarantees the citizen rights as against itself and other citizens. Again, this power accrues to him for himself because he is inherently important—not because he is son or father, peasant or overlord, member of a clan or a guild.

Doctrine of individualism

These ideas shift the emphasis of several thousand years of social beliefs and let loose innumerable implications. Individualism lowers the value of tradition and puts a premium on novelty. True, the individual soul had long been held unique and precious by Christian theology, but Christian society had not extended the doctrine to every man's mundane comings and goings. Nor were his practical rights and powers attached to him as a man but, rather, to his status within the social scheme. Now the human being as such was being officially considered self-contained and self-propelling; it was a new regime and its name was liberty.

Nature of the changes. The contents and implications of these potent words—liberty, equality, and fraternity, individualism and populism, simplicity and naturalness—aid in the understanding of the cultural situation of Europe at the dawn of the era under review. But these continuing ideas necessarily modified each other and in different times and countries were subject to still other influences.

For example, the active phase of the Revolution in France—say, 1789 to 1804—was influenced by the classical education of most of its public men. They had been brought up on Roman history and the tales of Plutarch's republican heroes, so that when catapulted into a republic of their own making, the symbols and myths of Rome were often their most natural means of expression. The eloquence of the successive national assemblies is full of Roman allusions. Later, when General Bonaparte let it be seen that he meant to rule France, he was denounced in the Chamber as a Caesar; when he succeeded, he took care to make himself first consul, flanked by two other consuls, to show that Caesar was in abeyance.

Changes in the fine arts

In the fine arts this Roman symbolism facilitated a thorough change of taste and technique. The former "grand style" of painting had been derived from royal and aristocratic elegance, and its allusions to the ancient classical past were gentle and distant, architectural and mythological. Now, under the leadership of the painter David, the great dramatic scenes of ancient history were portrayed in sharp, uncompromising outlines that struck the beholder as the utmost realism of the day.

In David's "Death of Socrates" and "Oath of the Horatii" civic and military courage are the respective subjects; in his pencil sketches of the victims of the Terror as they were led to execution, reportorial realism dominates; and, in his designs for the setting of huge popular festivals, David, in collaboration with the musicians Méhul and Grétry, provided the first examples of an art in scale with the new populism: the taste of the court and the city was transcended by the needs of the nation.

But it must be added that except for a few canvases and a few tunes (including the "Marseillaise") the quality of French Revolutionary art was not on a par with its aspirations. Literature in particular showed the limitations under which revolutionary artists must work: political doctrine takes precedence over truth, and the broad effects required to move the masses encourage banality. There is no French poetry in this period except the odes of Chénier, whom the Revolution promptly guillotined, as it did France's greatest scientist, Lavoisier. The French stage was flourishing but not with plays that can still be read. The Revolutionary playwrights only increased the dose of sentiment and melodrama that had characterized the stage at the close of the old regime. The aim was to hold up priests and kings to execration and to portray examples of superhuman courage and virtue. Modern opera goers who know the plot of Beethoven's *Fidelio* can judge from that sample what the French theatre of the Revolutionary years thrived on. Others can imagine for themselves Molière's *Misanthrope* rewritten so as to make Alceste a pure patriot and hero, undermined by the intrigues of the vile courtier Philinte.

It may seem odd that once the Revolution was under way there should be so much indignation and protest against courtiers, priests, and kings and such fulsome homage paid to virtue and patriotism. What accounts for it is the difficulty of transforming culture overnight. People have to be persuaded out of old habits—and must keep on persuading themselves. Even politically, the Revolution proceeded by phases and experienced regressions. Manners and customs themselves did not change uniformly, as one can see from portraits of Robespierre at the height of his power wearing a short wig and knee breeches, republican and Rousseauist though he was.

Napoleon's influence. After Bonaparte's coup d'etat tension eased as the high ideals dropped to a more workaday level, just as the puritanism was replaced by moral license. And, as a by-product of Bonaparte's expeditions to Egypt in 1798–99, a new stylistic influence was added to the Roman: sphinxes and the lion-claw motif appeared in furniture, and the Near East began to attract attention. The Roman idea itself shifted from republic to empire as the successful general and consul Bonaparte made himself into Napoleon.

Bonaparte's coup d'etat

The Emperor had an extraordinary capacity for attending to all things, and he was concerned that his regime should be distinguished in the arts. He accordingly gave them a sustained patronage such as a Revolutionary party rent by internal struggles could not provide. But Napoleon had tastes of his own, and he had to control public opinion besides. In literature (he had been a poet and scribbler in his youth), he relished the Celtic legends of *Ossian* and encouraged his official composer Lesueur in the composition of the opera *Ossian ou les Bardes.* In painting, he favoured the surviving David and the younger men Gros and Géricault, both "realists" concerned with perpetuating the colour and drama of imperial life, including the life of the battlefield. But to depict contemporary issues on the stage (except perhaps in the ballet, which was flourishing) did not prove possible, for the stage must present genuine moral conflict if it is to produce great works, and moral issues are not discussable under a political censorship.

The paradox of the Napoleonic period is that its most lasting cultural contributions were side effects and not the result of imperial intentions. Two of these contributions were books. One, Chateaubriand's *Génie du christianisme* (1802; *The Genius of Christianity*), was a long tract de-

signed to make the author's peace with the ruler and revigorate Catholic faith. The other, Madame de Staël's *De l'Allemagne,* was a description of the new culture in Germany. Napoleon prohibited the circulation of the book in France, but its message easily survived. The two other sources of future light were men. One was the group of philosophers known as Idéologues, who were scientific materialists and concerned with abnormal psychology. The other was Napoleon Bonaparte himself, or rather the figure of Napoleon as seen by his age after Waterloo.

General character of the Romantic movement. The mention of Waterloo (1815) suggests the need to make clear a number of chronological discrepancies. It has been possible so far to discuss the general shift in the temper of European life without naming fixed points. It sufficed to say "before or after 1789" or "from 1789 to the Napoleonic Empire." But from now on the generations of culture makers and the dates of some of their works must be duly situated, without on that account losing sight of unities and similarities in the onward march of artistic and intellectual movements. If, for example, one takes the Romantic poets, one finds that Goethe came to maturity in the 1770s, when the English Romantics were just beginning to be born. Their French, Italian, Russian, Polish, and Spanish counterparts were, in turn, born about the year 1800, when the English were already in mid-career. The same irregularity in the onset of Romanticism is found in the other arts, and it is complicated (at least superficially) by the names given to various movements and persons in the different countries of Europe. Thus, in Germany the term *Romantismus* is applied to only a small group of writers, and Goethe and Schiller are called classic. In Poland and in Russia, classic is likewise the label for the great writers whose characteristics in fact align them with the Romantics elsewhere.

Chronological discrepancies in the Romantic movement

All these accidents of birth and nomenclature can be taken in stride by remembering the patterns found in each country or decade and the reasons for their appearance at that time and place. Within the slightly more than half century between 1789 and 1848, the phenomenon of Romanticism occurred and developed its first phase. Those who made it may have come early or late, belonged to this or that nationality, proved to be originators or synthesizers of existing elements—all such considerations appertain to individual biography or the history of a particular art or nation. What matters in the evolution of European culture considered as a whole is the orchestration of all the voices as they come in to swell the ensemble.

The main purport of the Romantic movement is said to be a revolt against 18th-century Rationalism and a resulting variety of new attitudes and activities: a turning in upon the self, a love of nature, the rediscovery of the Middle Ages, the cult of art, a taste for the exotic, a return to religion, a fresh sense of history, a yearning for the infinite, a maudlin sentimentality, an overvaluing of emotion as such, a liberal outlook in politics, a conservative outlook, a reactionary outlook, a Socialist-utopian outlook, and several other "characteristic features."

It is clear that not all of these can be equally true, characteristic, or important, since some contradict the others. At the same time it was inevitable that so sweeping a cultural revolution as Romanticism should contain incompatible elements. For instance, the political opinions enumerated above did in fact win the allegiance of different groups among the Romantic artists and thinkers for a longer or shorter time. But—to take note of other supposed definitions—not all Romanticists returned to religion: Goethe and Berlioz were pantheists; Byron and Heine, atheists; and Victor Hugo, a sort of Swedenborgian. As for sentimentality, its occurrence was rather a hangover from the 18th century than a new fashion of feeling, for the Romantic cult of art and of strong emotion goes dead against the weak sentimental mood. Similarly, the taste for history, for the Middle Ages, and for the exotic shows a strong curiosity about the particulars of what is real though ignored by previous conventions. All critics, however, are agreed upon one Romantic trait: individualism. And it is here that the figure of Napoleon plays its cultural role.

Incompatible elements of Romanticism

Napoleon was regarded as the model of a new man, even during his life, because his career was manifestly the product of his own thought and will working against the greatest imaginable resistance. He typified the individual challenging the world and subduing it by his genius. A movement that numbered as many artists and geniuses as did Romanticism was bound to find in Napoleon the individual par excellence or, as might be said in modern jargon, a supremely autonomous personality. This perception explains why nearly all the great names of the first half of the 19th century are found on the roster of those who praised Napoleon—from Beethoven and Byron to Hazlitt and Stendhal and Manzoni. Some who were politically his enemies—Sir Walter Scott, for example—nonetheless respected and pondered over the miracle of his achievements. No comparable attention has been paid to the dictators of the 20th century, a fact sufficiently explained by the real difference between Napoleon and them. Stendhal, who had taken part in the Russian Campaign of 1812, analyzed that difference ahead of time: Napoleon had intellectual power—and not merely political and military. In whatever he did, he showed originality of conception, a stupendous grasp of detail in execution, and the utmost speed in acting out his vision. This sequence, translated to other realms, was the very pattern of the creative imagination. It also seemed the vindication of individualism as a philosophy of life: open the world to the individual and the world will witness marvels unimagined before.

These remarks about Napoleon should convey a sense of the Romantics' attitude toward themselves and their situation. It is true that culturally they stood in opposition to their immediate forebears. But all generations do the same, and yet it is only once in a while that the conflict produces great things. Romanticism had the paradoxical advantage of making its way during and after the 25 years of Revolutionary and Napoleonic wars. For the tumult of battle and political overturns did its share to clear the ground for artistic innovation. When habits and expectations are repeatedly upset and frustrated in the broad public realm, the general mind opens up to novelty offered in other realms. That is one avenue of cultural, stylistic, and emotional change. When Stendhal was expounding Romanticism to the French in 1822, he argued that to go on writing in the Neoclassic vein was "to provide literary pleasure for one's grandfather." His remark was readily understood—at least by his young readers. Mighty events had dug a chasm between past and present, making plain the remoteness of the 18th century.

The Romantics' attitude toward themselves

And yet a paradox remains. When the Romantic innovator produced his first characteristic work—say Wordsworth with the *Lyrical Ballads* of 1798—he had to wait a good while for a hearing though his sense of the death of Neoclassicism was clear; it arose from a poet's quick perception of the decay of forms. Already in 1783 Blake had written of contemporary English verse that "The sound is forced, the notes are few." But these two poets' estimate of the state of poetry was not widely shared till the passage of time—*i.e.,* of men, and events—had worn down old routines. It is generally said that Blake and Wordsworth and most other Romantic artists were ahead of their time; it would be more accurate to say "ahead of their audience."

This phenomenon, characteristic of the modern age, further complicates the question of when Romanticism emerged or was established in different countries. Even so, it is possible to assign, with sufficient reason, a particular decade when in a given nation the movement was in being and active in the production of important works, though not necessarily accepted by a large public. In England and Germany that decade was the 1790s: Blake, Wordsworth, and Coleridge; Goethe (with the first fragment of *Faust*), Schiller, Herder, Jean Paul (Richter), Beethoven, Tieck, Wackenroder, Hölderlin, Schelling, Schleiermacher, together with the beginning of a German translation of Shakespeare by Tieck and Schlegel, mark the advent of the new age.

In Italy, France, and Russia, the decisive years opened in 1820. They are signallized in Russia by the abundant poetic output of Pushkin, in Italy by the work of Manzoni and Leopardi and by the surrounding discussions of literary theory, and in France by the poems of Lamartine,

Vigny, Victor Hugo, and Mme Desbordes-Valmore. The paintings of Delacroix, the first compositions of Berlioz, and Balzac's *Chouans* show that a new spirit was at work. Finally, in the 1830s, Poland—through its poet and novelist Mickiewicz—and Spain—through the works of Rivas, Espronceda, José de Larra, and Zorrilla—joined the rest of Europe in its richest artistic flowering since the Renaissance: the leading nations can boast one or more Romantic artists of the first magnitude.

The fundamental Romantic purpose

Romanticism in literature and the arts. The fundamental Romantic purpose was to grasp and render the many kinds of experience that classicism had neglected or had stylized. Romanticism was the first upsurge of realism—exploratory and imaginative as to subject matter and inventive as to forms and techniques. The exploration of reality surveyed both the external world of peoples and places and the internal world of man. The Scottish and medieval novels of Sir Walter Scott, beginning with *Waverley* in 1814, illustrate the range of the new curiosity, for Scotland was a "wild" place, outside the centres of civilization, and the Middle Ages were similarly "barbarous" and distant in time. When Byron went to the Near East and used it as the setting of his adventure poems, he was creating the same sort of interest, this time in a region as real and "central" (to itself) as Paris and London. In both these writers, factual detail is essential to the new sort of effect: the scenery is observably true, and so is the history, given through local colour. As Byron said when criticized: "I don't care two lumps of sugar for my poetry, but my costume is correct." Blake, 20 years earlier, had taken a stand against Sir Joshua Reynold's academic doctrine that the highest form of painting depicted the broadest general truth. Said Blake: "To particularize is the only merit."

Particulars, moreover, are all equally proper for the artist; the use he makes of them is what matters. When Wordsworth and Coleridge sought to revivify English poetry, they hit upon two divergent kinds of subject: Coleridge took superstition and the folk tale and wrote "The Rime of the Ancient Mariner" in the form of an old ballad; Wordsworth took the modern street ballad—a kind of rhymed newspaper—and produced his versified incidents of common life in common speech. In France, where the division of the vocabulary into "noble" and "common" (*i.e.*, unfit for poetry) had been made and recorded in dictionaries, the Romantics led by Hugo used the prohibited words whenever they saw fit. Hugo's verse drama *Hernani* (1830) created a scandal in the audience when the heroine was heard to speak of her handkerchief and when a character did not use a roundabout phrase about the march of the hours to say: "It is midnight."

The rediscovery of Shakespeare

The importance of such details can hardly be exaggerated and can perhaps be best understood by recalling what the rediscovery of Shakespeare meant to the Romantics. His rise from grudging esteem, even in England, to European idolatry by 1830 had a significance beyond the one already mentioned of serving to put down French classical tragedy and, with it, French cultural tyranny. The German scholar, critic, and playwright Lessing was among the first to use Shakespeare for that purpose, but the arguments in his theatre reviews, called *Hamburgische Dramaturgie,* sprang from critical genius and not mere national resentment. Shakespeare spelled freedom from narrow conventions—the set verse form in couplets, the lofty language and long declamations, the adherence to verse throughout, the exclusion of low characters, comic effects, and violent action—or, in a word, from the atmosphere of the court and the aristocratic etiquette.

What Shakespeare typified (and not in poetry or plays alone—witness his influence on Berlioz) was the right of the creative artist to invent his own forms, loosen the joints of grammar and metric (or the canons of any art), follow the promptings of his spirit (tragic or gay, vulgar or mysterious, but in any case venturesome), and see where this emancipation from artificial rules led the muse. There was danger in freedom, as always; the conventions ensure safety. But the aim of the Romantic genius was not to play safe or even to succeed; it was to explore and invent, multiply modes of feeling and truth, and thereby revigorate a dead or dying culture. The motto was not common sense but courage. This resolve explains why the men who came to idolize Shakespeare also rediscovered Rabelais and Villon and revalued Spinoza, the lone dissenter who had worshipped a God pervading the cosmos; Benvenuto Cellini, the fearless artist at grips with the principalities and powers; and "Rameau's Nephew," the ambiguous hero of Diderot's posthumous dialogue, a strange figure disturbingly in touch with the dark forces of the creative unconscious.

Drama. With so much feeling astir and so many novel ideas being agitated, it might seem logical to expect a flourishing school of Romantic drama. But only a few isolated works, more interesting than irreplaceable, compose the dramatic output of the Romanticists—Shelley's *Cenci,* Byron's *Manfred,* and Kleist's brilliant pieces in several genres. Ironically, Shakespeare's new role as emancipator had a curiously paralyzing effect on the theatre down to the middle of the century and beyond. In England, poet after poet tried his hand at poetic drama, only to fail from stage fright at the thought of Shakespeare. On the Continent, various misconceptions about him and old habits of classical tragedy prevented a new drama from coming to life. Victor Hugo's plays contained brilliant verse, and their form influenced grand opera (Wagner's no less than Verdi's), but the fact remained: the dramatic quality could be found everywhere in Romanticist art except on the stage.

Reflection on this point suggests that, quite apart from Shakespeare, the very concern of the Romantics with exploring the inner and outer worlds precluded the writing of great plays. These perhaps require that one world or the other be taken as settled, at least at the start of the dramatic action. Be that as it may, the double reality in flux was the stimulus of all Romanticist endeavour.

Painting. This generality holds for the painters as well; their "reality," too, was by no means "given," so that the notation of fresh detail and the study of new means to transmute the visible into art occupied all those who came after David. Goya led the way in Spain by depicting the vulgarity of court figures and the horrors of the Peninsular War. In England, Constable painted country scenes with a vividness at first unacceptable to connoisseurs. He had to argue with his patron, Sir George Beaumont, about the actual colour of grass. To prove that it was not of the conventional brownish tint used by academicians, he seized a violin, ran out of the room with it, and laid it on the lawn, forcing the unaccustomed eye to perceive the difference between chlorophyll and old varnish. At the same time, Géricault astonished the Parisians by painting, in harrowing detail, "The Raft of the Medusa," not an antique and noble subject but a recent event: the survivors of a shipwreck adrift and starving on a raft.

The new "reality" in painting

The young Delacroix was emboldened by the example and, inspired also by the work of his English friend Bonington, began to paint contemporary scenes of vivid realism—*e.g.,* the Turkish massacre of the Greek peasants at Chios. Later, Delacroix was to visit Morocco (exoticism again) and to discover there the secret of coloured shadows and other pre-Impressionist techniques. His English counterpart, J.M.W. Turner, was pursuing the same goal of realistic truth, though along a different path that nonetheless also led to Impressionism—and beyond. When asked one day why he had pasted a scrap of black paper on a portion of his canvas, he replied that ordinary pigment was not black enough. And he added: "If I could find something even blacker, I would use that."

Sculpture and architecture. No similar transformations of the visual occurred in sculpture or architecture. Canova and Thorvaldsen continued to produce figures and busts on Neoclassical lines; and only Barye, the great sculptor of animals, and Rude, the creator of the Marseillaise panel on the Arc de Triomphe, showed any signs of the new passions. As for architecture, it may have been the love of history that prevented distinctive work. Pugin and Viollet-le-Duc did grasp the principles of what a new style should be, the former's love of Gothic reinstating the merit of framework construction and the latter's breadth of vision as a restorer leading him to predict that iron construction would one day pass from mere utility to high art.

It was actually in railway construction that the seeds of a new architecture were sown. Tunnels and bridges and terminals were needed as early as the mid-1830s, and unassuming engineers such as the Brunels and Robert Stephenson set to work to design them. All they had for solving the new and awkward problems of topography, speed, and cost were the ideas they drew from machinery and the vulgar materials they knew how to handle. The results were often remarkable, and they remained to inspire the makers of 20th-century steel and concrete architecture.

Music. It may seem as if the art of music by its nature would not lend itself to the exploration and expression of reality characteristic of Romanticism, but that is not so. True, music does not tell stories or paint pictures, but it stirs feelings and evokes moods, through both of which various kinds of reality can be rendered or expressed. It was in the Rationalist 18th century that musicians rather mechanically attempted to reproduce stories and subjects in sound. These literal renderings naturally failed, and the Romanticists profited from the error. Their discovery of new realms of experience proved communicable in the first place because they were in touch with the spirit of renovation, particularly through poetry. What Goethe meant to Beethoven and Berlioz and what German folk tales and contemporary lyricists meant to Weber, Schumann, and Schubert are familiar to all who are acquainted with the music of these men.

Changes in the relationship between music and literature

There is, of course, no way to demonstrate that Beethoven's *Egmont* music—or, indeed, its overture alone—corresponds to Goethe's drama and thereby enlarges the hearer's consciousness of it; but it cannot be an accident or an aberration that the greatest composers of the period employed the resources of their art for the creation of works expressly related to such lyrical and dramatic subjects. Similarly, the love of nature stirred Beethoven, Weber, and Berlioz, and here too the correspondence is felt and persuades the fit listener that his own experience is being expanded. The words of the creators themselves record this new comprehensiveness. Beethoven referred to his activity of mingled contemplation and composition as *dichten,* making a poem; and Berlioz tells in his *Memoires* of the impetus given to his genius by the revelation not only of Beethoven's and Weber's music but also of Goethe's and Shakespeare's works and—not least—of the spectacle of nature. Nor did the public that ultimately understood their works gainsay their claims.

It must be added that the Romantic musicians—including Chopin, Mendelssohn, Glinka, and Liszt—had at their disposal greatly improved instruments. The beginning of the 19th century produced the modern piano, of greater range and dynamics than heretofore, and made all wind instruments more exact and powerful by the use of keys and valves. The modern full orchestra was the result. Berlioz, whose classic treatise on instrumentation and orchestration helped to give it definitive form, was also the first to exploit its resources to the full, in the *Symphonie fantastique* of 1830. That work, besides its technical significance just mentioned, can also be regarded as uniting the characteristics of Romanticism in music: it is both lyrical and dramatic, and although it makes use of a "story," that use is not to describe the scenes but to connect them; its slow movement is a "nature poem" in the Beethovenian manner; the second, fourth, and fifth movements include "realistic" detail of the most vivid kind; and the opening one is an introspective reverie.

Self-analysis. In this Romantic investigation of the self, some critics have seen little more than excessive ego or, in modern terms, a tiresome Narcissism. No doubt certain Romantic works arouse boredom or disgust with hair-splitting analysis. But the boredom is often due to the fact that after a hundred years the discoveries have staled. When fresh, they came as a revelation; in the works of the great poets and novelists, in Hazlitt's essays and Jean Paul's fictions, and the irony of Byron's letters or Heine's journalism, the truth has not grown dim or platitudinous.

Desirability of self-analysis

It was in any case desirable that this extensive analysis of the self sould be attempted then, for only an age in which individualism was both theoretical and passionate could see the logic of the undertaking and act upon it. The logic was this: given the autonomous and unique individual, a search by himself into his moods, motives, fears, and loves must bring forth data otherwise unobtainable. Add these results together, and you have a repertory of clues to the inner life of mankind as a whole. For the uniqueness of each individual is bounded by traits he shares with his fellows, and this common element enables the psychologist to connect and organize the reports of the self-searchers. It is on this hypothesis, incidentally, that the demand for originality in art has continued unabated since the Romanticists. Forget the model, for there is no such thing; avoid conformity; discover your true self, the buried child; be authentic and sincere—these precepts, which still govern art and criticism, are the legacy of Romantic individualism.

Introspection naturally implies an inner life worth looking into, and most Romantic artists brought forth extraordinary findings. They form the groundwork of modern thought. One cannot easily imagine Freud or Joyce, much less the degree of self-consciousness shared by Western men today, without the deliverances of Blake, Wordsworth, Keats, Leopardi, Stendhal, Benjamin Constant, Sainte-Beuve, Heine, and innumerable other writers of the early 19th century. And towering above them as the creator of the prototype of Romantic introspection is Goethe with his Faust.

Faust was the figure in which a whole age recognized its mind and soul; and the adjective Faustian, as Spengler's use of it makes clear, still describes tendencies at work in culture today. The principal one, already mentioned, self-consciousness—the identity crisis—remains. The belief, moreover, that movement, activity, is better than repose and that striving is better than achieving is clearly the great postulate of contemporary civilization. Faust himself ends by giving his life to practical works in behalf of his fellow man. But he sets himself on that path only after a slow and deep analysis of his divided soul, which has been ruled together or in turn by despair, lust, superstition and the forces of the unconscious, the love of innocence, the conviction of sin and crime, the horrors of hypocrisy and conventional life, the temptations of wealth and power, the disgust with pedantry and established religion, the yearning for infinite knowledge, and through knowledge, peace. Faust, in short, traverses the whole cosmos, made up of the inner and outer worlds, to find in the act of self-dedication to humanity the justification of his existence.

Early 19th-century social and political thought. The Romantics who studied society through the novel or discoursed about it in essays and pamphlets were also devoted to the cause of humanity, but they arrived at politically different conclusions from Goethe's and from one another's. Scott and Disraeli were forerunners of Tory democracy as Burke was of liberal conservatism. Dickens, a passionate humanitarian, stirred the masses with his examples of the law's stupid cruelty, but he proposed no agency of betterment, content to despise Parliament, the law courts, and the complacency of the wealthy. Balzac wrote his "social zoology," *La Comédie humaine,* to show that society becomes a bloody jungle without the church and the monarchy to restrain human passions.

The cause of humanity

Stendhal noted the same reality but was more concerned with the free play of individual genius; he resigned himself to the social struggle, provided not too many stupid men ran the inevitably heavy-handed regimes. Freedom might be found by the happy few through the loopholes of a mixed government such as England's, whereas in the ostensibly free United States there was no protection against social pressure and no likelihood of genius, in art or in politics.

The great authority on American democracy was Tocqueville, whose astonishing survey in two volumes is still packed with useful lessons. Tocqueville confirmed Stendhal's low estimate of freedom of thought in America, but he foresaw in the United States the first example of a type of democracy that would surely overtake the Western world. He found in such a future many good things and many defects; he predicted a day when slavery would threaten disaster to America; he foretold what kind of poetry a democracy would produce and delineated the art of

Walt Whitman; he apprehended the complication of laws and the declining quality of justice; but he was reconciled to what must be.

Post-Revolutionary thinking. What lay behind all 19th-century writings on politics and society was the shadow of the French Revolution. In the 1790s the Revolution had aroused Burke to write his famous *Reflections* and Joseph de Maistre his *Considérations sur la France.* They differed on many points, but what both saw, like their successors, was that revolution was self-perpetuating. There is no way to stop it, because liberty and equality can be endlessly claimed by group after group that feels deprived or degraded. And the idea that these principles are universally applicable removes any braking power that national tradition or circumstance might afford.

Proof that the revolution marched on, slow or fast, could be read (as it still can be) in every issue of the daily paper since 1789. In the early 19th century the greatest pressure came from the liberals, whether students, bankers, manufacturers, or workmen enlisted in their cause. They wanted written constitutions, an extension of the suffrage, civil rights, and from time to time wars of national liberation or aggrandizement in the name of cultural and linguistic unity. For example, all the intellect of western Europe sided with Greece in the 1820s when it began its war of emancipation from Turkey. Byron himself died at Missolonghi while helping the Greeks. Poets wrote odes, and painters painted scenes of war. Between liberalism and national liberation the line could not be clearly drawn. In Italy, Germany, Poland, Russia, Spain, Portugal, and South America, revolt in the name of liberty was endemic until the middle of the century. Only England escaped, but it was by a narrow margin, after threats and violent incidents expressing the same animus as elsewhere.

Industrialization

Meanwhile, the first disturbances resulting from machine industry—sabotage, strikes, and conspiracies (for trade unions were generally held illegal)—reinforced the Revolutionary momentum, not only in fact but also in theory. As early as 1810 the business cycle, the doctrine of the exploitation of the worker, and the degradation of life in industrial societies had been noted and discussed. By 1825 the writings of the Comte de Saint-Simon, which proposed a reorganization of society to cure these evils, had won adherents; by 1830 the Saint-Simonians were an acknowledged party with sympathizers abroad, and by 1832 the words Socialism and Socialist were in use.

The Saint-Simonian doctrine relied on a benevolent dictatorship of industrialists and scientists to remove the inequities of the free-for-all liberal system. Other reformers, such as the practical Robert Owen, who organized successful communities in Scotland and the United States, depended on a strong leader using ad hoc methods. Still others, such as Leroux and Cabet, were Communists of divergent kinds seeking to carry out elaborate theories of the perfect state. Proudhon denounced the state, as such, and all private property. As a philosophical anarchist, he wished to substitute free association and contract for all legal compulsions. In England, the school of Bentham and Mill—Utilitarians or Philosophical Radicals—attacked existing institutions in the name of the greatest good for the greatest number, and by their arguments they succeeded in reforming the top-heavy legal system. Without doctrine but moved by a similar sense of wrong, Thomas Carlyle fought the Utilitarians for their materialistic expediency and himself sought light on the common problem by pondering the lessons of the French Revolution and publishing in 1837 what is still the greatest account of its catastrophic course. Later, Carlyle gave in *Past and Present* a suggestive picture of what he deemed a true community: quasi-medieval, based on the Faustian joy of work, and relying for its cohesion on its leader's genius and strength of soul.

Numerous other schemes of reform or revolt could be mentioned. One irony about them is that the name that has clung to most of them is utopian Socialism, which suggests unrealizable visions, whereas the historical fact is that a great many were tried out in practice, and some lasted for a considerable time. As in Carlyle's book, the force of character of one man (Owen was a striking example) usually proved to be the efficient cause of success.

Throughout this social theorizing, whatever the means or ends proposed, two assumptions hold: one is that men have a duty to change European society, to purge it of its evils; the other is that men *can* change society—they need only come together and decide what form the change shall take. These axioms by themselves, without the memory of 1789, were enough to keep alive in European culture the hope and the threat of continuing revolution.

The principle of evolution. Yet it should not be imagined that revolution by force inspired every thinking European. Even if liberals and reactionaries were still ready to take to the barricades to achieve their ends, the conservatives were not, except in self-defense. The conservative philosophy, stemming from Burke and reinforced by modern historical studies, maintained the contrary principle of evolution. Evolution indeed swayed as many 19th-century minds as its rival, and it was sometimes the same minds.

Evolution was the belief that lasting and beneficial change comes about by slow and small degrees. It is often imperceptible and therefore congenial to human habits. It breaks no heads and spills no blood; it is natural, organic. In truth, the idea of evolution is patterned on biology—the slow growth and decay of living things. More than that, evolution in the zoological sense of "descent with modification" had been a recognized speculation among men of science since 1750, when Buffon included it in his *Histoire naturelle.* Lamarck had elaborated the idea at the turn of the 18th century, while Erasmus Darwin, the grandfather of Charles, had by 1796 worked out for himself a compendious theory of similar import. In 1830–33 the geologist Lyell, setting forth the corresponding notion that changes in the Earth take place through the operation of constant and not cataclysmic causes, devoted a chapter to Lamarckian biology—to the evolution of species by imperceptible steps.

Revival of interest in history

As if these teachings were not enough to implant a form of thought, the revival of interest in history made easy and obvious the transition from the world of nature to that of man. It seemed logical to think of both as evolutions and even to liken the state to an organism. Certainly the student of institutions finds them steadily and profoundly altered by minute incidents and variations. Compared to these causes, the violent breaks made by war and revolution seem more superficial and less permanent.

The evolutionary scheme permitted several other beliefs while affording a number of arguments and conveniences. Anyone who had inherited from the previous era a faith in progress could now attach it to this motive power, evolution. Anyone who wished to classify nations or institutions by rank could place them as he thought proper on an evolutionary scale. Anyone who resisted change or wished to speed it up could be admonished with some evolutionary yardstick. Finally, anyone who intended to write a work of history or propaganda found his organizing principle ready-made for him. In the first half of the 19th century, every subject of interest, from costume to the criminal law, was presented in innumerable studies as proceeding majestically at an evolutionary pace.

Another way of stating the influence of this great idea is to say that the mind of Europe had experienced the "biological revolution." Whereas in the 17th century the Newtonian revolution had imposed the model of mechanics and mathematics, what impressed itself on the 19th century was the living organism—change and variety as against fixity and regularity. The logic of preferring "biology" to "mechanics" in an age of individualism, of realism about concrete particulars, and of passionate imagination and introspection need only be stated to be evident.

Science. This is not to say that the science of physics stood still during the Romanticist period. It was the time when the conservation of energy was established and the mechanical equivalent of heat demonstrated. There also prevailed the "physical" pseudo-science of phrenology, which professed to relate individual attributes to bumps and hollows in the skull, and which led to the physical anthropology that defined three, 10, 20, and 100 different races of man by the end of the century. Still, the 19th was more emphatically the century that furnished the theory of the cell (Schleiden and Schwann, 1838–39), which led

ultimately to the notion of microscopic creatures responsible for putrefaction and disease and, later still, to cytology and genetics.

It is noteworthy, too, that the 19th century saw the establishment of chemistry on the Daltonian hypothesis of the atom, but it was coloured by the "biological" notion of elective affinities to explain compounds. Goethe, who was an early evolutionist and the scientific expositor of the metamorphosis of plants, called his last novel of human love *Elective Affinities.*

Poetic ideals versus science

On the surface the poetic mind of the age seemed hostile to both science and technology. Wordsworth looks like an enemy of science when he says: "We murder to dissect" and deprecates the man who is willing to "peep and botanize upon his mother's grave." But reflection shows that the animus here is not so much against science as for the science of life. And to make the effort of recapturing the temper of the times is to be at once struck by the precarious status of science itself. Thought in the 18th century had ended in materialism and skepticism. Some writers, such as d'Holbach, had reduced all phenomena to the interaction of hard and unfeeling particles; others, such as David Hume, had "proved" that man can know nothing beyond his impressions and therefore can have no certainty about the truth of cause and effect, on which scientific statements depend. The Romanticist generations could neither agree that life was a concourse of unfeeling atoms nor trust the physicists' assertions based on a law of causation that the most acute thinkers discredited.

Such were the iron constraints within which the famous "crises of the soul" and conversions to religions new or old took place in the '20s and '30s of the last century. Carlyle, Mill, Lamennais, and many others described them in autobiographical works. The choice seemed to be a blind and meaningless universe or a meaning restricted to the motions of matter, equally blind. Even if the latter scheme "explained," it was vulnerable to David Hume's irrefutable doubts.

Early 19th-century philosophy. What enabled 19th-century culture to pursue the scientific quest and regain confidence in spiritual truth was the work of the German Idealistic philosophers, beginning with Immanuel Kant.

Kant. Kant took up Hume's challenge and showed that although we may never know "things as they are," we can know truthfully and reliably the data of experience. The reason for this certitude is that the mind imposes its categories of time and space and causation on the flowing stream and gives it shape. Science, therefore, is not a guess, nor is human knowledge a dream. Both are solid and verifiable. Indeed, certainty, according to Kant, extends as far as morals and aesthetics. The essence of morals is the commandment not to perform any act that one would not want to become a precedent for all human action and always to consider an individual as an end in himself, not as the instrument of another's purpose. The fusion in Kant of ideas stemming from Rousseau and the Enlightenment with ideas fitting the needs of the coming century (Kant died in 1804) made him the fountainhead of European philosophy for 50 years.

Kant's disciples. His disciples—Fichte, Hegel, Schopenhauer—twisted or amplified his teachings. Coleridge in England and Victor Cousin in France adapted to home use what seemed fitting. The school as a whole was known

German Idealism

as German Idealism because it relied on the distinction between the thinking subject and the perceived object; "idea" and "thing" were unlike, but idea (or the mind) played a role in shaping the reality of things, from which derived all stability and regularity in the universe.

Stability was desirable as a guarantor of natural science, but to any observer of human history it seemed contradicted by events, especially those since the French Revolution. Hegel, coming after Kant and witnessing Napoleon's victory at Jena in 1806, was therefore impelled to modify the Kantian logic. To account for the great torrent of human history, he imagined a logic of movement, by which opposing forces are in perpetual battle. Neither side wins, but the upshot of their struggle is an amalgam of their rival intentions. Hegel called the pros and the cons and their survivors thesis, antithesis, and synthesis. Human affairs are ever in dialectic (dialoguing) progression. At times a "world-historical figure" (Luther, Napoleon) embodies the aspirations of the masses and gives them effect through war, revolution, or religious reformation. But throughout the succession of events, what is taking place is the unfolding of Spirit or Idea taking on itself the concrete forms of the real. Hegel's was another version of evolution and progress, for he foretold the extension of liberty to all men as the fulfillment of history. It is interesting to note that until 1848 or '50 Hegel was generally considered a dangerous revolutionary, a believer in an irresistible progress that mankind must earn by blood and battle. Karl Marx, as a younger Hegelian, was to carry out Hegel's unspoken promise on a different base.

Other branches of the all-powerful German philosophy deserve attention but can be spoken of only as they relate to high Romantic themes. Fichte's modification of Kant made the ego the "creator" of the world, an extreme extension or generalization of individualism. At the other extreme, but more in tune with contemporary science and art, Schelling made nature the source of all energy, from which individual consciousness takes off to become the observer of the universe. Nature is a work of art and man is, so to say, its critic, and because human consciousness results from an act of self-limitation, it perceives moral duty and feels the need to worship.

Religion and its alternatives. That need made itself felt ecumenically throughout Europe from the beginning of the 19th century. It had indeed been prepared by the writings of Rousseau as early as 1762 and the pastoral work of John and Charles Wesley much earlier. The atheism and materialism of the late 18th century was in truth far more responsible for the religious revival of the 1800s than the brief secularism of the French Revolution itself, so that when the Catholic writings of Chateaubriand and Lamennais in France, the neo-Catholic Tractarian movement in England, and the writings of Schleiermacher and his followers in Germany began to take effect, their success was due to the same conditions that made Romanticist art, German idealism, and all the "biological" analogies succeed: the great thirst caused by dry abstractions in the Age of Reason needed quenching. Religious fervour, artistic passion, and "gothic" systems of philosophy filled a void created by the previous simple and mechanical formulas.

The political consequences of the religious revivals

The religious revivals, Catholic or Protestant, also aimed at political ends. Their participants feared the continuation in the 19th century of secularism and wholly material plans. In every country the liberals proposed to set up in the name of tolerance ("indifference," said the Christian believers) governments that would serve exclusively practical (indeed commercial) interests. Church and state were to be separated, education was to be secular, which would really mean anti-religious. National traditions would be broken, forgotten, and youth would grow into "economic man," Benthamite Utilitarian man, with no intuition of unseen realities, no sensitivity to art or nature, no humility, and no inbred morals or sanction for their dictates.

Scientific Positivism. The alternatives to the richness of a life lived within traditional religion were two: scientific Positivism and the cult of art. The name Positivism is the creation of Auguste Comte, a French thinker of mathematical cast of mind, who in 1824 began to supply a philosophy of the natural sciences opposed to all metaphysics. Science, according to Comte, delivers unshakable truth by limiting itself to the statement of relations among phenomena. It does not explain but describes—and that is all mankind needs to know. From the physical sciences rise the social and mental sciences in regular gradation (Comte coined the word sociology), and from these man will learn, in time, how to live in society.

Having elaborated this austere system, Comte discovered the softer emotions through a woman's love, and he amended his scheme to provide a "religion of humanity" with the worship of secular saints, under a political arrangement that the sympathetic John Stuart Mill nonetheless described as "the government of a beleaguered town." Comte did not attract many orthodox disciples, but the influence of his Positivism was very great down to recent times. Not alone in Europe but also in South America it

formed a certain type of mind that survives to this day among scientists and engineers.

The cult of art. The second "religious" alternative, the cult of art, has had even greater potency, being at the present time the main outlet for spirituality among Western intellectuals. In the Romantic period this fervour was allied with the love of nature and the idolatrous admiration of the man of genius, beginning with Napoleon. A writer as sober as Scott, a thinker as cogent as Hegel, and an artist as skeptical as Berlioz could all say that to them art and its masters were a religion; and they were not alone. At the death of Goethe in 1832, Heine inveighed against the great man's followers who made art the only reality. In the second and third Romantic generations, born around 1820, the religion of art grew still more pronounced and took on an antisocial tone that became more and more emphatic as time passed. "Art for art's sake" ended by signifying, among other things, "art the judge of society and the state." This doctrine was expounded in full detail by the Romantic poet Gautier as early as 1835 in the preface to his entertaining and sexually daring novel, *Mademoiselle de Maupin.* In those pages the familiar argument against bourgeois philistinism, against practical utility, against the prevailing dullness, ugliness, and wrongness of daily life was set forth with much wit and that spirited defiance of all established things which is associated rather with the 1890s and the present day than with Romantic passion. Its occurrence then is but another proof that Romanticism was the comprehensive culture from which later styles, thoughts, and isms have sprung—down to our own time.

"Art for art's sake"

The middle 19th century. During the half century when Romanticism was deploying its talents and ideas, the political minds inside or outside Romanticist culture were engaged in the effort to settle—each party or group or theory in its own way—the legacy of 1789. There were at least half a dozen great issues claiming attention and arousing passion. One was the fulfillment of the Revolutionary promise to give all Europe political liberty—the vote for all, a free press, a parliament, and a written constitution. Between 1815 and 1848 many outbreaks occurred for this cause. Steadily successful in France and England, they were put down in central and eastern Europe under the repressive system of Metternich.

A second issue was the maintenance of the territorial arrangements of the treaties made at the Congress of Vienna in 1815. Metternich's spies and generals also worked to keep this part of the post-Napoleonic world intact; that is, the boundaries that often linked (or separated) national groups in order to buttress dynastic interests. Except in Belgium, the surge of national, as distinct from liberal, aspirations throughout Europe was unsuccessful in the 1830s. But defeats only strengthened resolve, particularly in Germany and Italy, where the repeated invasions by the French during the Revolutionary period had led to reforms and stimulated alike royal and popular ambitions. In these two regions, liberalism and nationalism merged into one unceasing agitation that involved not merely the politically militant but the intellectual elite. Poets and musicians, students and lawyers joined with journalists, artisans, and good bourgeois in open or secret societies working for independence: they were all patriots and all more or less imbued with a Romanticist regard for the people as the originator of the living culture, which the nation was to enshrine and protect.

Merger of liberalism and nationalism

To be sure, this patriotic union of hearts did not mean agreement on the details of future political states, and the same disunion existed to the west, in England and France, where liberals, only half satisfied by the compromises of 1830 and 1832, felt the push of new radical demands: of the Socialists, Communists, and anarchists. Reinforcing these pressures was the unrest caused by industrialization—the workingman's claims on society, expressed in strikes, trade unions, or (in England) Chartist groups. This cluster of parties agitated for a change that went well beyond what the advanced liberals themselves had not yet won. Add to these movements those that purposed to hold still or to restore former systems of monarchy, religion, or aristocracy, and it is not hard to understand why the great revolutionary furnace of 1848–52 was a catastrophe for European culture. The four years of war, exile, deportation, betrayals, coups d'etat, and summary executions shattered not only lives and regimes but also the heart and will of the survivors. The hoped-for evolution of culture was broken and, with all other hopes and imaginings, rendered ridiculous. The search began for new ways to achieve, on the one side, stability and, on the opposite, the final desperate revolution that would usher in the good society.

For although they seemed decisive, the battles of '48 and after did not, in fact, test the worth of any one idea. Nationalism won and lost in different parts of Europe. Liberalism gained in Italy and Switzerland, but was set back in Germany and France. English Chartism seemed to collapse, yet its demands began to be carried out. The Socialist experiment in France (Louis Blanc's national workshops) also seemed discredited; yet the ensuing regime of Napoleon III made attempts, however clumsy, to deal with poverty by welfare methods. Repression had shifted men and territories—Metternich was gone from Austria—but Germany and France were not on that account flourishing under liberty, equality, and fraternity. There was peace, but war was imminent; and subversive groups continued to plot and publish and frighten the bourgeois, while industry and urbanization contributed their gains at the cost of the now familiar miseries and sordor.

In these circumstances the mind of Europe suffered an eclipse, followed by a protracted mood of despondency. Many established or emerging artists and thinkers had been killed or torn from their homes or deprived of their livelihood. Richard Wagner fleeing Dresden, where he conducted the opera; Chopin and Berlioz at loose ends in London, because in Paris music other than opera was moribund; Verdi going back to Milan with high patriotic hopes and returning to Paris in a few months, utterly disillusioned; and Victor Hugo in exile in Belgium and later in Guernsey—all typify the vicissitudes in which men of reputation found themselves in mid-career. For the young and unknown, such as the poet Baudelaire or the English painters who formed the Pre-Raphaelite Brotherhood, it was no time to invite the public to admire boldness and accept innovation. Critics and public alike were all nerves and hostility to subversion. To read Flaubert's masterpiece, *L'Éducation sentimentale,* is to understand the atmosphere in which the first phase of Romanticism ended and its ramified sequels came into being.

Mood of despondency

Realism and Realpolitik. The dominant feeling was that high hopes had perished in gunfire, and this realization bred the thought that hope itself was an error. Any new effort must therefore stay close to the possible, the "real." Realism with a capital *R* and *Realpolitik* together sink their roots in a distrust of man's imagination. This grim caution born of harsh experience coincided with a sense of fatigue that made Romanticist work seem like the foolishness of youth.

The appropriate cultural note must no longer be the infinite or heroic or colourful but rather their opposites. If the commonly accepted term Realism for this reaction of the 1850s is used, it must be with these presuppositions in mind. For the Romantic passion for the particular and exact was a realism, too; it was what Dr. Johnson much earlier had called "vehement real life." The Realism of the disillusioned '50s dropped the vehement, the passionate and, in order to run no risk of further illusion, limited what it called real to what could be readily seen and felt: the commonplace, the normal, the workaday, and often the sordid.

Limitation of "real"

In the same spirit *Realpolitik* rejected principles. The word did not mean "real" in the English sense; in German it connotes "things"—hence a politics of adaptation to existing facts, pursuing plain objects, admitting no obligation to ideals. In this light we can understand the unexpected epithet "scientific" that Marx and his followers bestowed on their brand of Socialism. It was a science not merely because it was presumably based on the laws of history but even more because in its view the advent of the Socialist state was to result from the interaction of things (classes, means of production, and economic necessity)

and not, as in earlier socialism, from the will (that is, the imaginative efforts of thinking men). The "objective" appearance given to the new politics of things, Socialist or other, generated that tough, no-nonsense atmosphere, which men then wanted as a source of reassurance in all their dealings.

Scientific Materialism. This search for certainty went with a swinging back of the pendulum in science itself from the vitalism of the previous period to the Materialism of the midcentury. German philosophers derided Idealism and taught the equivalence of consciousness and chemistry: "without phosphorus, no thinking." The machine once more became the great model of thought and analogy—and nowhere more vividly and persuasively than in biology, where Darwin's advocacy of natural selection won the day because it provided a mechanical means for the march of evolution. The struggle for life (Spencer's phrase of 1850, adopted by Darwin in his subtitle) obviously had the requisite "toughness" to convince and, like *Realpolitik,* it followed no principle—whoever survived survived. That Darwin to the very last included other factors in his theory of evolution—Lamarckian "use and disuse" as well as direct environmental forces—carried no weight with a generation bent upon machine certainty. These secondary explanations were ignored, in the usual way of cultural single-mindedness, and for 30 years after the publication of the *Origin of Species* in 1859, an orthodoxy of universal mechanism reigned over all departments of thought.

It prevented the recognition of Mendel's work on genetics; it put religious, philosophical, and ethical thought on the defensive—only what was "positive" (*i.e.,* material) held a presumption of being real and true. The same reasoning produced a school of social Darwinists who saw war between nations and economic struggle among men as beneficent competition leading to the survival of "favoured races"—another phrase from Darwin's subtitle.

The "really real"

And by a final twist of logic, Materialism reinforced the moral gloom of the period by casting doubt on both the permanence and the validity of all that was being redefined as "really real." For on the one side, the second law of thermodynamics guaranteed the cooling of the Sun and the pulverization of the cosmos into cold and motionless bits of matter; and, on the other, orthodox "machineism" brought its leading prophets, Huxley and Tyndall, to consider men and animals as automatons moved as helplessly as atoms and planets. Consciousness is an epiphenomenon—in plain word, an illusion—precisely as in Karl Marx consciousness and culture are illusions floating above the reality of economic relations.

Victorian morality. To be sure, not everybody in Europe believed or cared about these affirmations. And although ideas long debated do in the end filter down to the least intellectual layers of the population, the time and place of triumph for a philosophy are limited by this cultural lag—a fortunate delay, without which whole societies might collapse soon after the publication of a single book. What kept mid-19th-century civilization whole was a subdued faith in the reality of all the things Realism and materialistic science denied: religious belief, civic and social habits, the dogma of moral responsibility, and the hope that consciousness and will did exist.

The sum of these invisible forces is conveniently known as the Victorian ethos or morality, a formula applicable to the Continent as well as Britain and one whose meaning antedates not only the midcentury revolutions but also the accession of Queen Victoria in 1837. Like Romanticism, this powerful moralism had its roots in the late 18th century—in Wesleyan Methodism, the Evangelical movement, in Rousseau, Schiller, and Kant. Its earnestness was of popular origin; it was anti-aristocratic in manners, and it sought the good and the true in a simple, direct, unhesitating way. Perceiving with warm feeling that all men are brothers under God, the moral man saw that slavery was wrong; and having so concluded, he proceeded to have it abolished by act of Parliament (Britain, 1833).

Such fervent convictions when widely shared exert tremendous power, and this concentration of belief and emotion made Victorian morality long impregnable. As Chesterton said of the Victorian painter Watts:

> He has the one great certainty which marks off all the great Victorians from those who have come after them: he may not be certain that he is successful, or certain that he is great, or certain that he is good, or certain that he is capable: but he is certain that he is right.

The sense of rightness generated a sense of power, which the Victorians applied to the monumental task of keeping order in a post-revolutionary society.

Partly by taking thought and partly by instinct, they perceived that the drive to revolution and the sexual urge were somehow linked. Therefore they repressed sexuality; that is, repressed it in themselves and their literature, while containing it within specified limits in society. Further, they knew that the successful working of the vast industrial machine required a strict, inhuman discipline. The idolatry of respectability was the answer to natural waywardness. To pay one's bills, wear dark clothes, stifle individual fancy, go to church regularly, and turn aggression upon oneself in the form of worry about salvation became the approved common modes of pursuing the pilgrimage of life.

Idolatry of respectability

It could not be expected that everybody would or could conform. From its beginning to the end, the Victorian age numbered a galaxy of dissenters and critics who scorned the conformity, called the religion a sham, and viewed respectability as mere hypocrisy. But the front held, and the massed forces behind it were at their strongest after the multiplied assaults of 1848.

Nothing gives a better idea of the astonishing moral structure called Victorianism than the development of the London Metropolitan Police, begun under Sir Robert Peel in 1829. A lawyer and a former captain who had fought in the Peninsular War were the first joint commissioners and creators of the force. At first they had to weed out the drunks and the bullies who had been the main types of recruit in earlier attempts at policing cities. At first, too, the people both ridiculed and fought with the new police. Gradually, the "peelers" came to be trusted; they remained unarmed regardless of circumstances; they learned to handle rioters without shedding blood; and in the putting down of crime they finally enlisted the public on their side. For something less than a century this unique relationship lasted, in which "law-abiding" and "police" were terms of respect—correlative terms, since the peelers (later "bobbies") could not have become what they were without the self-discipline and moral cohesion of the "respectable."

The upheavals of the midcentury, cultural as well as political, put Victorianism to a severe test, for after wars and civil disorders laxity is natural, and ensuing despair induces a reckless fatalism. There was cause indeed for apprehension. When the Great Exhibition of 1851 was planned on a scale hitherto unattempted, many expressed the fear that to allow tens of thousands from all over Europe to come together under the Crystal Palace was to invite massive riots. Ministers and heads of state would be assassinated. In the event, no protracted assembly of common people and their leaders was ever so quiet and orderly. The moral machinery worked as efficiently as that which was on display under the glass dome.

The advance of democracy. Yet, while a stringent moralism held in check endemic subversion and anarchy, Darwinism and the machine analogy stimulated endless forms of self-consciousness. If man could fashion and continually improve these engines, perhaps he could also engineer an improved society. Because evolution was at last "proved," thanks to Darwin, perhaps it also gave warrant for social and political progress by gradual steps. Spencer's all-inclusive philosophy, likened then to Aristotle's, foresaw an inevitable movement from the simple and undifferentiated to the complex and specialized—as in modern life. Clearly, whether automatons or not, men kept thinking and having purposes; and among evolutionists and scientific socialists alike, thought and purpose included the hastening by voluntary action of what was sure to come by force of natural laws. These and other desires acting in the light of Realism and taking shape in the increasing organization of the toiling masses brought Europe to accept democracy as inevitable.

Evolution as a warrant for social and political progress

The word democracy is used here in a cultural sense. It does not imply a set of political institutions so much as the signs and the agencies that herald the coming populist state of our day: for example, the extension of the franchise, in parliamentary or plebiscite form; the secret ballot; the legalization of trade unions; the rise of a Catholic social movement; the passage of education acts providing free, public, and compulsory schooling; the formulation of Tory democracy as a cure for the evils of free-for-all economic liberalism; the beginnings of welfare legislation (in France under Napoleon III, in Germany under Bismarck); the secularization of life by state action, by the prestige of science, and also by the liberal movements within the churches themselves; and finally, after a decade or so of public education, the great extension and popularization of the press. At the passage of the Reform Act of 1867 in Britain, which gave the vote to urban workingmen, Robert Lowe had said, "Now we must educate our masters." In a parliamentary system the means to that education cannot be the schools alone. The adult "common man" must continually be informed and appealed to for his satisfaction as well as for coherent policy in government. The instrument for this purpose was the new journalism. The quarterlies of the early 19th century gave way to the monthlies in the '60s and they in turn to the weeklies, while the daily papers, costing now but a penny and simplifying all they touched, began to reach the millions.

Realism in the arts and philosophy. In the period of so-called Realism, the arts and philosophy as usual gave both form and substance to the prevailing fears and desires. The mood of soberness and objectivity was alone acceptable, and what art presented to the public confirmed the reasonableness of the mood.

Literature. This interaction accounts for such things as the marked change of tone in Dickens' novels that occurs between *David Copperfield* (1850) and *Bleak House* (1853). The temper expressed in most concentrated form the very next year in *Hard Times* now dominates Dickens' mind and works to the end: life is a dreary sort of underworld; happy endings are artificial conventions not to be believed.

The same mood explains why Flaubert's *Madame Bovary,* which ranks today as the realistic novel par excellence, and is on all counts grim enough in its rendering of boredom and vulgar misery, was judged "too artistic" by some contemporary critics, not close enough to the most common of realities, that of common speech.

Yet the evolution of Flaubert's work remains instructive for an understanding of Realism as a literary creed. Flaubert had begun by writing a highly coloured, imaginative story on *La Tentation de Saint Antoine,* which the author's friends advised him to burn, tone down, or rewrite. Flaubert put it aside and began the novel that became *Madame Bovary.* Its setting was the provincial world around him, not the Egyptian desert; the characters were of the most ordinary type, not an improbable Christian ascetic haunted by visions. But even in the working out of his plain tale Flaubert had to subdue his lyrical Romantic genius to the discipline he had adopted. The description of a rainstorm, for instance, had to be done over and over again so that it would not stand out and be "interesting" by virtue of the observer's mind. It had to be made ordinary and the observer kept outside, just as in science. *Madame Bovary,* begun as a magazine serial, was soon censored by the editor and then prosecuted as immoral by the state. For Flaubert's Realism had gone so far as to portray in no flattering colours the dreary lives and motives of average provincials of both sexes, and the picture violated the rules of the indispensable moralism. What is more, the fate of Flaubert's unhappy heroine symbolized what had happened to the more daring and poetic-glorious time before 1848: as Flaubert said, Emma Bovary was himself.

The significance of *Madame Bovary*

His novel is thus simultaneously a model and a critique of the new genre—a critique, too, of the state of Europe that produced it. Many other writers between 1850 and 1890 pursued matter-of-factness without this ulterior effect and rendered the details of middling life with such impassiveness and fidelity that to this day many use "realistic" as a synonym for dreary or sordid and regard "the novel" as a reliable historical source. On the precise definition of Realism, George Gissing gave, through a character in one of his own novels, a brilliant commentary: the character is at work on a novel which shall be so true to the dullness of daily life that no one will be able to read it.

Painting and sculpture. The term Realism applies no less to the plastic arts than to literature, but in painting and sculpture it proved difficult to give form overnight to the change of attitude just noticed in literature and political life. The transition between the passionate poetry and drama of Géricault and Delacroix and the Realism of Courbet and Manet was gradual. It came by way of the "open-air" school of Barbizon, whose landscapes seemed arid (at least to the classically trained academic painters of the day) and pointless in the sense that they depicted the commonplace. Still, when the full shock of Realism inflicted by the works of Courbet and Manet occurred, it was severe: here were coarseness and violence in manner and subject. Courbet's backgrounds are thick and his people drab; Manet's nude "Olympia" is no goddess nor even a beautiful woman; she is a prostitute, and her name seems like a piece of irony. The portrait of his parents is a painful representation of simple poverty unrelieved by any glow of spirit or intelligence—yet the work itself is beautiful: such was, throughout, the aim and achievement of Realism.

The gradual transition to Realism in painting

In England, by an historical accident, pictorial realism was embodied in subjects that seem far removed from the commonplace. The school that took up the challenge against academic painting and modified the vision of Constable and Turner called itself Pre-Raphaelite. Its members were Holman Hunt, John Millais, and Dante Gabriel Rossetti, and the name they took for their "brotherhood" expressed their resolve to paint like the masters who came before the imitators of Raphael. It is necessary to put it in this clumsy way in order to make clear that Raphael himself was not being condemned, but only his academic followers who introduced "unreality."

To be a Pre-Raphaelite was to see the world with a sharp eye and an undistorting mind and to render it with intense application to solidity of form, bright colour, and natural pose and grouping. All this was to be understood from the motto *Death to Slosh!* In order to make the new virtues vividly clear and also because the Pre-Raphaelites were reared on great literature, their subjects tended to draw upon legend, or Dante, or the New Testament. It was the conception and treatment that constituted the innovation. Everybody could see it, because it went against the habit of "pretty-pretty" illustration. In fact the nominal subject dropped out of sight in the startled response to form and colour. Paradoxically, then, the commonplace subjects of the French realists and the legendary ones of the English Pre-Raphaelites were alike insignificant when compared with the effort to re-create by art the texture and "feel" of actuality—and nothing more. Such was precisely the goal Flaubert pursued and reached in *Madame Bovary.*

Popular art. It hardly needs to be added that this conscious purpose of high art could interest but a relatively small portion of the public and that, for the growing mass of readers of fiction and viewers of art, other kinds of satisfaction were necessary. The ordinary three-volume novel from the lending library and the continued serial in the magazine or newspaper supplied the demand by aping, adapting, and diluting not one but half a dozen literary tendencies, old and new. The number of novels produced in all languages in the 19th century has never been estimated, but it surely must be of the order of astronomical magnitudes. And the whole output was realistic in the sense that it professed to impart the real truth about life. It was contemporary in setting and speech, took the form of a history, and taught its readers how other people lived. The pictorial counterpart was the "chromo," the cheap colour lithograph that illustrated either fiction or news stories in forms which, however false they must seem to a critical eye, again gave the illusion of commonplace reality.

Growth of the novel in the 19th century

Music. At first sight, it would seem as if music were a medium in its nature resistant to Realism. But that is

to reckon without the obvious use that music has always made of sounds directly associated with life—church bells, hunting horns, military bands, and the like. In an age when Realism was at a premium, the opera would be the form where these and other associations easily found their place. So it was in midcentury Europe, where Meyerbeer and others provided the effects to suit the fussily "real" staging of all plays, musical or not. Clocks, tables, animals, waterfalls, and especially costume could be relied on to be genuine up to the limit of the possible: live bullets for real deaths were shied away from, and real lightning was out of reach.

But a genius who is often mistakenly grouped with the Romantics, Richard Wagner, supplied this ultimate deficiency—and by musical means. As critics have pointed out, Wagner's system of leitmotives, or musical tags that denoted an object, a person, or an idea, was consciously or unconsciously an accommodation of Realist intent to operatic understanding. This is true not simply because the musical notes "wave" up and down as Isolde waves her scarf at Tristan—a trivial enough device of a sort found in many composers; it is also true in the deeper sense, which constitutes Wagner's unique genius, namely that he was able to compose great music that was steadily and precisely denotative of items in the story by repeating and interweaving their assigned musical tags. (J.Ba.)

The modern age

CHRONOLOGY OF THE MODERN AGE

During the 19th century economic change had an overwhelming impact on international affairs. At the opening of the century, Britain had been the foremost commercial state in the world, and British wealth had been the mainstay of the coalitions against Napoleon. Britain had been the first nation to undergo the Industrial Revolution, and its early lead in this field, combined with its far-flung empire, ensured its primacy of power in Europe and the world for the rest of the century. Britain was also dedicated to the policy of free trade, adopted in the 1840s.

After 1870 Britain retained its primacy, but the margin of its advantage diminished, challenged in the main from three quarters—the United States, Japan, and Germany. But the impact of the first two is essentially a 20th-century story; the German, on the other hand, was quite clear, sharpened before 1914 by the German policy of naval construction; thus Germany, after unification, lay at the centre of international politics.

The European record over 100 years seemed to belie the gloomy predictions of overpopulation made by Thomas Malthus at the beginning of the century. Although some 50,000,000 Europeans swarmed to other continents, the population of Europe roughly trebled. But the condition of the masses generally improved. The fundamental reason for the change, allowing for political agitation and some humanitarian pressure, lay in the increase in the production of wealth, industrialization most of all. The 19th-century experience of Europe on this score constitutes a perhaps exceptionally fortunate accident, for industrial Europe could draw on the resources of the outer world for food and raw materials.

This growing wealth had other repercussions besides that of raising the standard of living. The combined impact of the rights of man and of scientific and technological development was favourable to the rise of the masses, one aspect of which was the spread of literacy, education in general, and a cheap press. The business of government became less and less the appanage of a restricted privileged class, though the conduct of foreign affairs continued to be a largely esoteric preserve. Yet, at election time and as the franchise spread, the temptation could not be resisted to appeal to the masses. The obverse facet of the democratic coin is demagoguery. How to conduct foreign policy in a democratic milieu has been an increasingly delicate and complicated problem. In the age of nationalism, the simplest way to arouse the masses was to appeal to their national prejudices and emotions. The enthusiasm with which the peoples of Europe responded to the outbreak of World War I in August 1914 bears out the point.

Outside of the United States and Japan, Europe was the mistress and the powerhouse of the planet. The state of European culture, both in the scientific domain and in the letters and the arts, was added grist to the mill of those who confidently believed that Europe *was* civilization, even if this civilization was an increasingly materialistic, godless one. As to the proliferation of arms, it was used as an argument for confidence in the lastingness of peace, ensured by what is now called the balance of terror, although the phrase itself is of later coinage.

There were Cassandras to be sure, and in the end they were, at least in a sense, to be proved right; but, at the time, they were mainly voices crying in the wilderness. When they went to war in 1914, the peoples and the governments of Europe had little inkling that they were initiating the closing chapter of the European Age.

The Bismarckian period, 1870–90. *Bismarck's system of alliances.* Otto von Bismarck, the German chancellor, was a practical man rather than one dedicated to abstract principle. His social bent was undoubtedly conservative, but he understood the active forces of the modern world. Having achieved his purpose of Prussian enhancement in the form of uniting Germany under the Hohenzollern crown, he was now satisfied and entertained no further aggressive designs. He believed that the maintenance of peace and the status quo would provide the best climate in which the new Germany could thrive. Bismarck's peace policy

During the 20 years that he remained in charge, his skillful diplomacy dominated the international scene.

Germany had no enemies save France, which had been humiliated in 1870 and stripped of Alsace-Lorraine. Bismarck had no fear of France if France could be kept in isolation. In France he preferred the republican government and even favoured colonial ambitions, which he hoped would divert French attention away from the Rhine and revanche while possibly embroiling France with others. As for the British, Bismarck was content to leave to them their monopoly of empire.

His relations with Russia were good, and in the new hyphenated Austria-Hungary the Hungarian half was satisfied with the *Ausgleich* (Compromise of 1867) and would emphasize a compensating interest toward the southeast—the Balkans. This would, however, give added point to possible Austro-Russian rivalry. Bismarck's own solution for this problem was simple: divide the Balkan apple of discord between the two contenders—he would be Austrian in Serbia and Russian in Bulgaria. Moreover, the three great conservative states of Europe shared an outlook reminiscent of the Metternichian.

These thoughts bore fruit in 1873 in the Dreikaiserbund (League of the Three Emperors). But agreement was not long lasting, for by 1876 local Balkan events resulted in a falling out between Austria and Russia. Bismarck's control had limits. If he must choose between the two, though he would rather not, he would opt for the wholly European and partly Germanic Dual Monarchy. The Austro-German Alliance was concluded in 1879, henceforth the cornerstone of German foreign policy. In Bismarck's eyes it was undoubtedly a defensive alliance, designed as much to reassure as to restrain Austria.

Within two years Bismarck was able to revive the tripartite connection, but the second Dreikaiserbund foundered, like the first, over local Balkan differences—this time over a war between Bulgaria and Serbia in which Serbia was rescued by Austrian diplomatic intervention.

This drove Bismarck into an extremely complicated scheme. His alliance with Austria was a fixed point, and he even informed the Russians of its terms. But in order to insure against the nightmare of the war on two fronts—a Franco-Russian coalition—he contrived an understanding between himself and Russia, the Reinsurance Treaty of 1887. This treaty has been described by some as Bismarck's greatest diplomatic achievement, by others as an attempt at political bigamy. It was a tightrope-walking act and lasted only as long as Bismarck himself was in power. The Austro-Serbian Alliance and the connection with Romania rounded out the Bismarckian network of eastern alliances. The Reinsurance Treaty

Italy had a low power rating in Bismarck's eyes, but, in

order to secure Austria's rear and to prevent a possible Franco-Italian connection, he enlisted Italy as well in his camp. The Triple Alliance of 1882, made possible in part by Italian fears of French designs, implied a stabilization of the still-existing Austro-Italian antagonism (a legacy of the Risorgimento) and a muting of the Italian irredentist claim against Austria. Italy also fitted into Bismarck's desire for general stability in the south.

Balkan problems. Ottoman affairs, particularly in the Balkans, which were still in the process of achieving national emancipation, continued to be a sensitive and unstable area. In the mid-1870s, revolts flared up in Bosnia and in Bulgaria, out of which ensued a Russo-Turkish war in 1877–78. The victorious Russians imposed the Treaty of San Stefano on the Turks. The treaty called for a drastic rearrangement of the Balkans, whereby a large Bulgaria would be created under the protection of Russia. But the Russian gains in the treaty alarmed the powers. In the face of determined British opposition, the Russians therefore agreed to a collective European revision of the treaty. The powers met at the Congress of Berlin in 1878 and produced a new Balkan arrangement: a much smaller Bulgaria was created and a series of compensations arranged for some of the powers, both large and small. The Balkan provinces of Bosnia and Hercegovina were put under Austrian occupation and administration, and Cyprus was "temporarily" occupied by the British. The Treaty of Berlin was a source of dissatisfaction to the Russians, who, after all, had won the war, and the Dreikaiserbund was a victim.

Seven years later, the Balkans flared up again, though the fighting, Serbo-Bulgarian, remained confined to their locale, and Bulgaria, joined by Eastern Rumelia, achieved full independence from the Sultan. As mentioned above, these episodes dissolved the second Dreikaiserbund.

The new imperialism. Events in the more peripheral regions of the Ottoman Empire are perhaps better placed within the purview of colonial activity. At Berlin in 1878 it had been suggested to the French, by both the Germans and the British, that Tunisia might be suitable compensation for them. Italy was also interested in Tunis, where the reality of an Ottoman connection was cloudy.

In France there were those who insisted that revanche and the German danger should be the first cares of French policy, while a colonial party thought this a sterile attitude that merely prevented France from taking part in Europe's growing interest in the wider world. In 1881 the latter group scored a success with the establishment of a protectorate over Tunisia. Italy was annoyed and joined the Triple Alliance, but there were no repercussions in Constantinople.

In the 1870s both the French and British became deeply involved in Egypt. The French interest was of long standing, and the Suez Canal was mainly used by British shipping. The Khedive's finances were mismanaged and the British and the French intervened to impose a joint Anglo-French supervision of Egyptian finances. Local Egyptian conditions led to further Anglo-French interference, and in 1882 the British took military action, the French having declined at the last moment to join them. This was the beginning of the effective British control of Egypt, for the supposedly temporary occupation of the country proved to be but the initiation of a continuing and deepening involvement.

Anglo-French involvement in Egypt

The cases of Tunis and Egypt may be conveniently regarded as the launching of the new imperialism, a vigorous recrudescence of Europe's imperial activity. The longer established British and French priority in the imperial domain was being emulated by others. Even initially anti-colonial Bismarck yielded to the pressure generated by the rapidity of German economic growth; by the mid-1880s Germany claimed various areas in Africa, and even Italy was developing an interest around and in Abyssinia.

In modern garb the age of discovery and conquest had returned, and the partition of Africa was under way. With a view to avoiding conflict, acknowledging the equal right of all, the powers met in Berlin in an attempt to put some order in the process. Actually, the Berlin West Africa Conference (1884–85), the result of a Franco-German initiative, affirmed the freedom of navigation and trade in the basins of the Congo and Niger rivers. The Free State of the Congo, the outcome of a private initiative headed by King Leopold II of the Belgians, achieved international recognition, and general rules for the establishment of colonial claims were agreed upon.

Penetration of Asia

There were manifestations of imperial activity (primarily British, Russian, and French) in Asia as well. The penetration of China was continuing, as was Russian expansion in Central Asia. This last, pushing toward the western and northern back doors of India, was a source of concern to the British, who found themselves involved in protective, "defensive" expansion. The two influences met in Afghanistan. Likewise, on the opposite side of India, a British protectorate over Burma was proclaimed in 1885, while the French became involved in war with China as a result of their activity in Tonkin.

This colonial activity was largely conducted by a handful of men with small bodies of troops. The peoples of Europe to a large extent were unaware of this aspect of the activity of foreign and colonial offices and had in general no consciousness of being involved in war. While there were also critics of imperialism among Europeans, theirs was but a small voice.

The conscious leaders of Europe's imperial urge nurtured ambitious schemes. Granting the priority of the economic interest and the rationalization of the white man's burden, or *mission civilisatrice,* the factor of national interest, of prestige and pride, should not be minimized. Control of Egypt meant control of the vital artery of the Nile. It was not long before the British were involved in The Sudan, where they suffered a temporary setback. They were also active at the southern tip of the African continent, where Cecil Rhodes, businessman and statesman of ruthless vision, was at work. From these two points as bases, Cairo and the Cape, plus British East Africa, the vision of a vast empire, roughly the eastern half of Africa, symbolized by the Cape-to-Cairo railway project, was born.

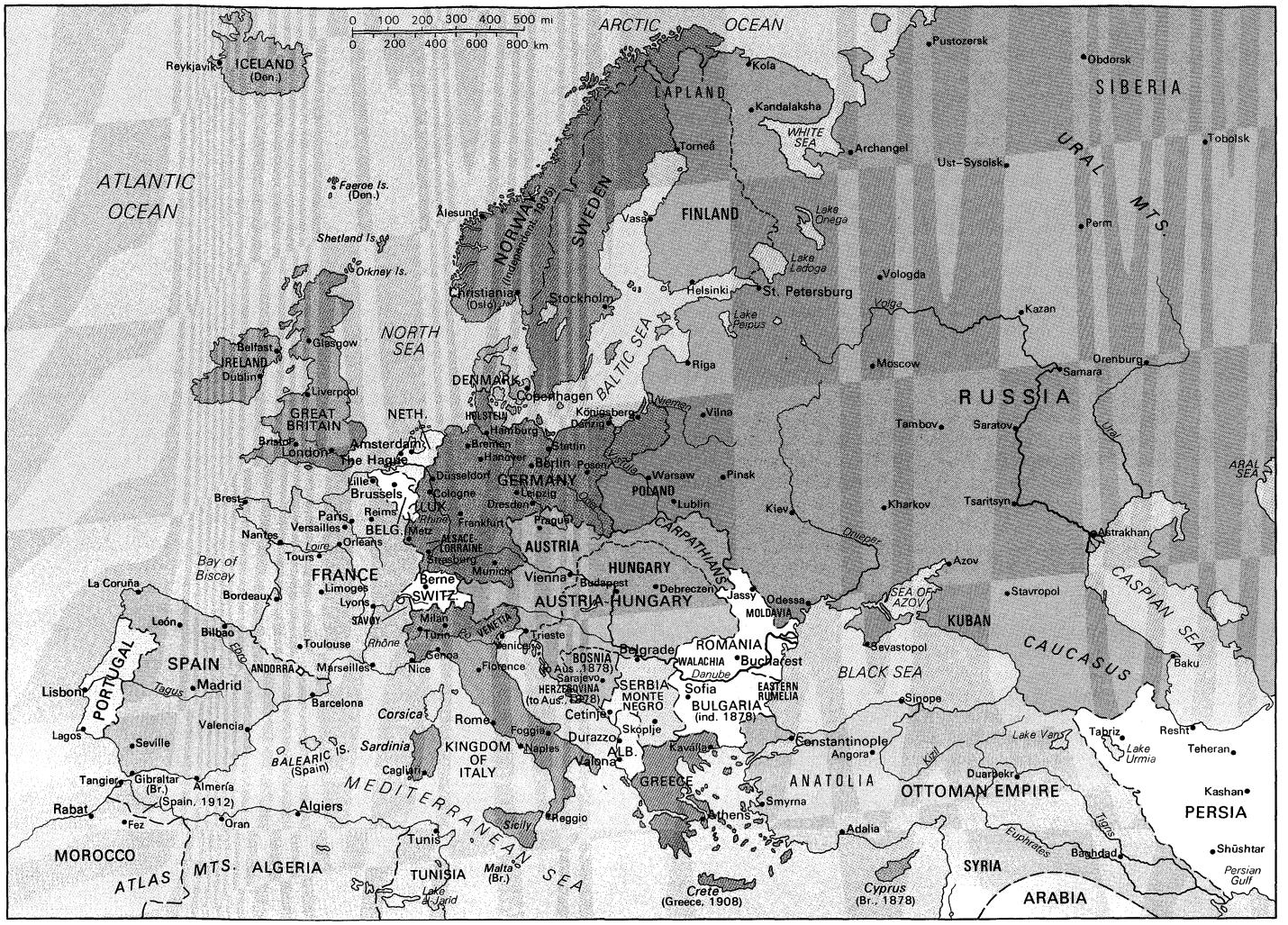
The French could entertain a comparable dream, an extension of France across the Mediterranean all the way to the Congo. Around the sources of the Nile, British, French, German, and Belgian influences met, to which the Italian may be added. Here was incipient rivalry and a potential source of conflicts, but these did not flare up until the 1890s, when exploring activity had laid sufficient bases for overlapping political claims.

Meanwhile, the focus of international relations, over which Bismarck presided, was still predominantly in Europe. But already the seeds of disintegration of the Bismarckian system were planted. The combination of the Triple Alliance, his own treaty with Russia, and the Mediterranean agreements constituted a fragile edifice, undermined by inner inconsistencies. Russia and France were the two revisionist states that Britain was willing, to a degree at least, to help contain. But difficulties, of an economic nature at first and not necessarily irretrievable, began to appear in the Russo-German relationship. Bismarck's dismissal in March 1890 opened a new chapter in the relations among the European powers.

The German challenge. *The period of transition, 1890–1904.* It is a proper measure of the authoritarian nature of the German state that Bismarck could be dismissed from office by the emperor; no constitutional issue was at stake. Bismarck's own place in the roster of statesmen is secure; if personally not a very attractive character, his rank nevertheless belongs among the first: like a good craftsman, he used the proper tool for the right job at the right time; war having served his purpose, he endeavoured to maintain the peace, all the while directing the energies released by the very fact of his unifying success.

Assessment of Bismarck

One reason for his dismissal was the advent, in 1888, of a new emperor, William II. Friction developed between the crusty old man and his young, arrogant, touchy, and romantic new master. The last Hohenzollern proved to be, in a sense, the most adequate and, for that very reason, the most unfortunate embodiment of the insecure, arrogant success that was Wilhelmine Germany, all too effectively conveying to others an impression of aggressive intent out of proportion to the reality of that intent.



Europe, 1871–1914.

A major foreign-policy decision faced the new German administration—whether or not to renew the Reinsurance Treaty with Russia. There were solid arguments pro and con, and the Tsar was desirous of renewal; but the German answer was finally refusal. This decision compounded the uneasiness that had already begun to cloud Russo-German relations before 1890, one consequence of which was the fact that Paris took the place of Berlin in the financing of Russia.

France had been anxious to escape from the isolation in which Bismarck had successfully kept it confined. But France and Russia had few common interests. Russia felt as little concern for French revanche as France did for Russia's Central Asian expansion, with its consequence of Anglo-Russian friction. The focus of France's rivalry was also Britain, to be sure, but it lay mainly in Africa, where Russia had no interest. The only possible binder of a Franco-Russian connection was common fear of Germany, which in the Russian case could arise from the Austro-German alliance.

The Franco-Russian connection began with a series of loans, launched on the Paris market, from 1888 on. A Russian purchase of French arms in 1889 was the next step, followed by military conversations, the elaboration of a contingency plan. The Tsar felt little attraction for what he considered a godless republic, but, eventually, in 1893, he underwrote the results of the military discussions, while the French, on their side, evaded their own constitutional issue through the contention that no formal political alliance was involved, thereby bypassing parliament. Yet from 1893 it is proper to speak of a Franco-Russian alliance as existing. This was the first major breach in the Bismarckian system, made possible in part by the confident but incorrect German assumption that such a possibility was excluded. Like all such instruments at this time it was defensive, and the terms of it were secret.

Franco-Russian alliance

But the situation was still very fluid, and the focus of European rivalries was becoming increasingly extra-European. The Russians were pleased when, in 1893, friction developed between France and Britain in Southeast Asia. The difference was composed, however, and Siam, as a buffer, retained its independence.

By this time the Japanese had made sufficient progress in their imitation of European ways to have developed imperial ambitions of their own. They turned to China, with which they dealt successfully in war, imposing upon it their demands in the Treaty of Shimonoseki in 1895. This Japanese success was of concern to Russia, and an unexpected combination ensued, for Germany—this was its way to weaken or undo the newly formed Franco-Russian connection—gave Russia support in opposing Japan, while France, rather embarrassedly, followed suit. The tripartite combination, Britain excluded, was successful in frustrating Japan of some of its gains; it had special appeal to the German Kaiser, who again and again returned to his project of a continental league.

Africa was at this time also the locale of a similar imperial competition. When in 1894 the British arranged for the purchase of a strip of the Belgian Congo that would connect their African holdings in the north and the south—the Cape-to-Cairo scheme—joint Franco-German pressure caused them to abandon the plan.

The Fashoda incident

But, on the entire course of the Nile, Britain would not allow any control other than its own. When in 1895 the French launched from Gabon an expedition led by Capt. Jean-Baptiste Marchand that was to march across Africa, effecting a connection with their East African base of Djibouti, a cautionary warning was issued in London. But in

September 1898 Captain Marchand reached the Nile at Fashoda, near where British forces under Lord Kitchener had been conducting operations. The confrontation of two handfuls of men on the Nile grew into a major clash between Paris and London.

It was a delicate passage. In Paris the French foreign minister, Théophile Delcassé, a convinced colonialist, came to the conclusion that war with Britain was not desirable. In the face of British intransigence, there was nothing to do but yield. French feeling ran high, but French prestige was seriously injured. Yet out of this situation was to develop one of the most important changes in European power relationships.

Britain's interest in South Africa had been enhanced by the discovery of gold and diamonds in the Transvaal, and Britain was thus brought into collision with that little Boer republic and its twin, the Orange Free State. The South African War began in 1899, after the Fashoda incident had been settled. The British underestimated the guerrilla capacity of the Boers. As a consequence, the protracted conflict entailed a far greater effort on Britain's part than had been thought necessary. In the end the Boers could not stand up unassisted against the resources of Britain. In 1902 the Boer republics were annexed into the colony of the Cape.

The episode had important repercussions. In Britain it led to soul-searching, arousing both strong criticism and nationalistic emotions. Of greater significance was its impact on Britain's international position. The worldwide sympathy for the Boers drove home to Britain the fact of its diplomatic isolation. Britain's traditional stance, that of preferring to remain uncommitted and to eschew alliances in peacetime, might stand in need of reconsideration in the light of the range of Britain's worldwide commitments and of challenges from a variety of quarters. Britain's margin of advantage was diminishing in the face of other rising powers.

Three powers—the United States, Germany, and Japan—were growing at a faster rate than others but at the moment it was Germany that mattered. Since the traditional rivalries with France and Russia still existed, the possibility of a German one being added was a proper source for concern in Britain, especially if it should develop into the Kaiser's plan of a continental league. Preservation of the divisions of Europe, the balance of power, had long been the first directing rule in the defense of Britain's interest.

The German challenge had two aspects. The first was commericial, an expression of the faster German economic growth. It could hardly be expected that Germany should deliberately restrain its expansion in that domain. The other, not unrelated to the first, was in essence political, reflecting the German desire for recognition by others of its right to a place in the sun. Such expressions as *Weltpolitik* and "our future lies on the water," fondly espoused by the Kaiser, expressed it to perfection. In concrete terms, it took the form, in 1898, of embarking on the building of a large navy; it was a considered decision, taken in full awareness of its political implications, the credit—or blame—for which was shared by the Emperor himself and Adm. Alfred von Tirpitz, the father of the German navy. It could hardly be expected that Britain would remain indifferent to such a decision.

Thus, the British reaction to the shifting relationships of power was to re-examine the merits of isolation; put in different form, to reduce the range of commitments by coming to terms with one of the rivals. Bearing in mind Fashoda and the rivalry with Russia across the breadth of Asia, not to mention the currently popular view of the racial affinity and superiority of the Nordic peoples and the tradition of friendly relations with Prussia, Germany was the logical candidate for an approach.

Joseph Chamberlain, a convinced imperialist, opened negotiations with the Germans. Discussions took place on three occasions between 1899 and 1901 but resulted in failure to reach an understanding, a failure that was in large measure due to German psychological misjudgment. The advice of the German chancellor, Bernhard, Fürst von Bülow, to the Kaiser to hold the bid high, regarding Britain as a demandant from whom a political commitment of abstention from continental involvement could be extracted, had a counterproducing effect. Proud Britain may have had difficulties but, after all, had had its way with both the French and the Boers. Even Chamberlain, a Britisher above all in the last analysis, had cautioned his German interlocutors that there were possibilities other than Germany open to Britain.

The first possibility materialized in an unexpected quarter. In 1902 an Anglo-Japanese alliance was concluded that was not even secret. Though the alliance represented a break in the traditional British avoidance of peacetime alliances, it did lighten the burden of British responsibilities in the Far East and thus enhanced Britain's freedom of action in Europe.

The opening years of the century witnessed a marked shift in the power relationships of Europe. Following nearly a decade of unfriendly relations with France, Italy put an end to the tariff war with the conclusion of a commercial treaty; it also reached an understanding with France over Tunis. Francesco Crispi's dream of an East African empire was frustrated by the defeat of Adowa in 1896. In 1900 Italy further acknowledged France's interest in Morocco in exchange for a reciprocal French acceptance of Italian hopes in Tripoli. The ambiguous Interpretation of the Triple Alliance that Italy gave France in 1902 to a considerable degree voided that pact of content, though it was simultaneously renewed. Bülow urbanely turned a good face on Italy's potential infidelity, but Germany entertained few illusions on the score of Italian dependability; Italy's position has aptly been described as being "on the fence."

Franco-Italian relations

In Serbia a domestic upset resulted in the murder of the pro-Austrian king, an Obrenović, and his replacement by the representative of the rival Karageorgević house; henceforth, Serbia abandoned its subservient Austrian allegiance and turned to Russia for protection.

The most important diplomatic shift at the opening of the century, however, was in Anglo-French relations. Fashoda convinced Delcassé that France should eliminate its traditional rivalry with England. He was ably assisted in this policy by Paul Cambon, his ambassador in London, who found receptive ears in Britain, especially after Chamberlain's failure with Germany.

Unlike the Germans, the French did not insist on a political commitment from England. Protracted Anglo-French discussions resulted in the Entente Cordiale of 1904, whereby France agreed to abandon its withering Egyptian hopes, Britain, in turn, sanctioned French ambitions in Morocco, and a number of other important differences were settled. In the context of the imperial activity of the day, this was a fair and reasonable quid pro quo. The understanding contained no hint of an alliance; it was a public document, but secret clauses were attached to it.

The meaning of this in 1904 was problematic, and it might have proved to be no more than a passing flirtation. Time and succeeding events proved it otherwise, however, and it became the opening of a significant new chapter in the affairs of Europe. The British and the French continued to differ on many things, but what gave solidity to the Entente was the fundamental fact that, allowing for differences of power in 1904, the two countries were increasingly finding themselves placed in a similar defensive position vis-à-vis the rising powers. With a time lag of roughly 50 years, the relative decline of the French power position was followed by a comparable British decline. Here was not a passing condition but rather a long-term change in the position of both countries.

The last decade of peace, 1904–14. Two fundamental assumptions of German policy—the impossibility of either a Franco-Russian or an Anglo-French connection—had proved mistaken. The precise nature of the second was uncertain, but Bülow soon had the opportunity to clarify its meaning. The proximity of two dates is significant: in February 1904 the Japanese, without warning, attacked and destroyed a Russian squadron in Port Arthur; in April the Entente agreement was concluded. France's ally, Russia, was at war with Britain's ally, Japan. The Anglo-Japanese alliance had a restraining effect on the possibility of French assistance to Russia; Britain and France shared

The Russo-Japanese War

a common interest in Russo-Japanese accommodation, and the Far Eastern conflict remained localized.

To the surprise of much of the world, the "little yellow men" inflicted a series of defeats on a major European power. To be sure, Russia was handicapped by logistics—the difficulty of supplying a theatre of war over a still incomplete Trans-Siberian Railroad—while the Japanese enjoyed the converse advantage of proximity, hence, of shorter lines of communication. They controlled the sea as well; the pathetic odyssey of Russia's Baltic fleet, sailing halfway around the world, ended in its destruction in the Battle of Tsushima Strait (May 27, 1905). There was nothing for it but to acknowledge defeat, all the more so as internal Russian conditions were taking on a revolutionary hue in 1905. The interposition of American good offices brought the belligerents together at Portsmouth (New Hampshire), where Japan collected the fruits of its competent preparations, having passed the test that insured great-power status. Manchuria and Morocco are far apart, but the simultaneity of the two situations created a connection between them. Britain could indeed renounce its own claims in Morocco; it could certainly not dispose of the rights of third parties. In the form of an ostentatious visit that the German kaiser paid to the Sultan of Morocco, Germany asserted its claim to an equal position in Morocco. The French tactics had been to isolate Germany, and this had been one of their purposes in reaching understandings with Britain, Italy, and Spain in regard to Morocco. On both sides of the Rhine there was complete understanding of the facts and the realities of the Moroccan situation; while Germany felt that Morocco logically "belonged" to France, it would claim a price, as others had, for its consent to that outcome. The issue was, precisely what price? The highest possible was the German view; the smallest was the French.

But Morocco was also a pretext. By asserting equal rights in Morocco, Germany could discover the depth of Britain's commitment in support of France. Bülow himself described his stance as that of "the Sphinx": keep others in uncertainty, and frighten the French; either or both of two results might thereby be achieved: a wrecking of the budding Anglo-French Entente, or the securing of a high price for Morocco.

The Moroccan crises

This is the substance of the First Moroccan Crisis, which meant on the German side raising the spectre of possible war. The French were indeed frightened, to the extent that Delcassé found himself alone among his Cabinet colleagues in opposing a policy of accommodation, insisting that Bülow's tactics were bluff. Properly, he resigned. In what had come to be a duel between two men, the German chancellor had scored the major victory of forcing the resignation of the French foreign minister. But, unwisely, Bülow sought to push his advantage too far, for, if the French yielded to the German demand for an international conference, they were resentful of what they considered undue foreign interference in their domestic affairs.

The Conference of Algeciras in 1906 was a surprise to Germany, which found itself virtually isolated, and the resulting Act of Algeciras, in effect, enhanced the French position in Morocco. Bülow had been all too successful in conveying abroad the impression of German aggressive intent, implemented by tactics of intimidation, the mailed fist, and sabre rattling. One result of the episode was the strengthening of the Entente, the opposite of his intentions; Anglo-French military discussions, again a contingency plan, began in 1906.

After a few years of quiescence, the Moroccan issue flared up again when the French moved in to restore order in the country. The Agadir Crisis of 1911 bears close resemblance in its unfolding to the one just outlined. Sending a German gunboat (the "Panther") to Agadir was a way of asserting a German claim equal to that of the French. From the very outset of the crisis, Britain took a much stronger position in support of France than it had in 1905, and after a tense passage the issue was finally resolved as it might have been in the first place, through a bilateral Franco-German understanding: Morocco to France in exchange for a substantial section of the French Congo to Germany. But the manner in which that result had been procured left a legacy of bitterness and suspicion, especially in Franco-German relations; in France there soon followed the *réveil national,* a resurgence of nationalistic feeling. Between the two Moroccan crises a good deal else had happened.

The impact of the impression created by German tactics in the First Moroccan Crisis and the continued growth of the German navy are to be largely credited for another development. Britain would maintain its naval superiority but proceeded further to reduce the range of its commitments. The Anglo-Russian agreement of August 1907 brought into existence the Triple Entente. The Anglo-Russian understanding is comparable to the Anglo-French one of 1904, dealing with Asiatic rivalries in this case. The heart of it was a division of Persia into three zones—British, Russian, and neutralized. The cry went up in Germany of *Einkreisung,* and indeed encirclement it was, though not with aggressive intent but rather as a response to German impatience expressed in tactics that to others seemed aggressive. Appearance can count for more than reality, and a vicious circle of misunderstanding was closing.

German encirclement

The forces of nationalism, meantime, continued to threaten both the Ottoman and the Habsburg empires. The policy of the latter oscillated between the grant of belated and insufficient concessions and repression. The South Slavs in particular were a troublesome problem, which the shifting of Serbia into the Russian camp in 1903 served to accentuate.

In 1897 Austria and Russia had together "put the Balkans on ice," but the Russian defeat in the Russo-Japanese War shifted the focus of Russian policy back toward Europe. In 1906 the Russian and the Austrian foreign ministers, Aleksandr Izvolsky and Aloys Lexa von Aehrenthal, met and agreed on a scheme for certain changes in the arrangements of 1878. Where Austria was concerned, it would put an end to the ambiguity of its position in Bosnia-Hercegovina by outright annexation of these provinces. The looseness of the understanding and the uncertainties arising from a successful Young Turk takeover in Constantinople—formal title to Bosnia-Hercegovina was still Turkish—induced Austria to create the *fait accompli* of annexation in October. This was the opening gun of the Bosnian annexation crisis of that year and the following. Russia this time, as Germany had done in the case of Morocco, demanded a European conference, which Austria, like France in 1905, was seeking to avoid. The tension lasted until the following spring, when Russia, militarily unprepared and but lukewarmly supported by France, felt that it could not face the prospect of hostilities when Germany took a determined stand in support of its own ally.

The outcome was an undoubted success for the Austro-German combination—a preview in many respects of July 1914—emphasized by a Serbian surrender to an Austrian ultimatum that demanded formal acknowledgment of satisfaction with the annexation. Germany had been the key factor in the resolution of the crisis, but it had acted in support of an initiative that was Austrian; the Bismarckian direction of the alliance had been reversed, in part because Germany considered that in the existing constellation of powers it could not afford to allow its one dependable ally to suffer a setback. But, as a consequence, the distrust and resentment of which Germany was the focus had been deepened.

There followed a two-year lull in the accelerating tempo of crises, during which, however, the arms race was intensified. In 1912 Viscount Haldane, the British war minister, visited Germany in yet another attempt to moderate the naval competition. The effort foundered on the usual reef of the German demand for a political commitment on England's part. Meanwhile, the Agadir Crisis had again raised the spectre of war. Though also resolved, as has been indicated, that confrontation may be regarded as the opening of the last chapter in the story of nearly a half-century of peace.

The Franco-German agreement had not even been concluded when Italy declared war on Turkey for the purpose of acquiring the vilayets of Tripoli and Cyrenaica. The Italians' diplomatic preparations had been adequate, all

the powers having at various times acknowledged the validity of the Italian claim. With some difficulty Italy had its way, and the war was concluded with the surrender of Libya by Turkey in October 1912.

The Italo-Turkish War and the Balkan Wars

The Italo-Turkish War was unpopular with the powers, who feared that an attack on the Ottoman Empire might lead to larger complications. These fears were wholly justified. In March 1912, the Russians scored a notable diplomatic success in contriving a Serbo-Bulgarian alliance, which was joined by Greece within two months. It was clearly an offensive alliance, and in October the Balkan allies went to war against Turkey. A final attempt by Austria and Russia to restrain them came too late.

The allies were surprisingly successful. By December, save for a few isolated centres of resistance, the Turks were defeated and surrendered the entirety of their Balkan holdings with the exception of their capital and the straits connecting the Mediterranean with the Black Sea.

The prearranged division of the spoils, underwritten by Russia, proved unacceptable to some of the other powers. Austria, supported by Italy in this case, interposed a veto on Serbian access to the Adriatic. The Balkan allies fell out. Bulgaria, unwisely, attacked the Serbs and the Greeks and found itself confronted with the Turks and the Romanians as well and shortly went down in defeat. The Treaty of Bucharest (August 1913) settled matters for Bulgaria, but in the western Balkans the powers imposed their own will by creating the state of Albania.

This settlement may be considered as the last instance of the successful operation of the Concert of Europe. The outcome of the Balkan wars was, nevertheless, regarded on the whole as a setback for the Austro-German combination, and the specific form of the settlement had the effect of exacerbating Serbian feeling toward Austria, now the exclusive focus of South Slav irredentism.

Outbreak of World War I

The explosion of the powder keg of Europe, the Balkans, in 1912–13 had failed to ignite the whole continent. But it was only six months before Austria decided to give its own South Slavs a demonstration of its power and determination. On June 28, 1914, the heir to the Habsburg crown, Archduke Francis Ferdinand, visited Sarajevo, the capital of Bosnia, to review army manoeuvres. Some young and irresponsible nationalists, idealistic martyrs to a cause in another view, succeeded in carrying out their plot. Francis Ferdinand and his wife fell victims to their pistol shots—shots that, more than an earlier one, truly were destined to ring around the world. The opening week of August saw all the major powers of Europe, save one, and two minor ones at war. The last chapter of the European Age had begun. (R.A.-C.)

World War I and the Russian Revolution. The war that began in 1914 was the first war of the masses, the first between nation-states able to command the energies of all their subjects and the resources of modern industrial technology, the first on a scale large enough to disrupt the world economy that now hinged on Europe. It differed completely from the wars of the second half of the 19th century; these had been engagements limited in purpose and scope, whereas the War of the Grand Alliances was as complex in purpose as it was unlimited in commitment. It was unprecedented in European history. (See INTERNATIONAL RELATIONS.)

Russia, France, and Great Britain being allies, Germany was from the start confronted with a war on two fronts. Its inevitable purpose was to crush France before Russian resources could be fully brought to action in the east. Its initial effort to achieve this by means of the Schlieffen Plan failed in 1914. Thereafter, fighting on the western front became a war of attrition, with millions on both sides locked in close and indecisive stuggles extravagant in men and materials. Artillery and the machine gun made each attack so costly in lives and munitions that soon each side had to recruit mass armies and expand munitions production on a scale neither had anticipated. On the eastern fronts warfare was more mobile but hardly less expensive in human life.

The strain on national resources, financial and industrial, but even more on discipline and morale, was immense. In the end it shattered the cohesion of the weaker dynastic empires of Turkey, Austria-Hungary, and Russia, and even the more solid Western states experienced mutinies and political crises. Yet the most remarkable feature of the war was the amount of hardship and sacrifice modern society could endure without disintegrating. The German submarine war on British, Allied, and eventually neutral shipping and the Franco-British blockade of the central powers imposed on civilian populations severe shortages of food and supplies. For all these reasons a new concept of total war emerged. Governments controlled or commandeered supplies, trade, credit, transport, and manpower more fully than ever before. The war accustomed nations to centralized power and widespread governmental activities. Total war made possible the 20th-century conception of the totalitarian state. Walter Rathenau, the industrialist who organized the German economy for war, was the forerunner of Lenin, who made Russia, after 1917, the first of the single-party totalitarian states.

The Turkish question. The first war problem confronting the Entente Powers was to clarify their alliance. On september 5, 1914, they signed the Declaration of London, by which they agreed not to conclude separate peace. The Allies second problem was to secure the neutrality of Turkey. The German-Turkish alliance was still unknown to the Allies. The Turks used the negotiations with the Entente to complete their mobilization. Complications arose when two German cruisers that had fled into the Dardanelles were bought by the Turks but continued to be manned by German crews. On October 29 these German warships, together with the Turkish fleet, bombarded Odessa. Russia, Britain, and France declared war on Turkey early in November. The Sultan proclaimed a "holy war" but could not subvert the Muslim subjects of the Allies.

The Dardanelles campaign

Because Turkey's entry into the war closed Russia's main supply line, the Western Allies decided to open the Dardanelles by force. This posed a ticklish problem, since a successful campaign would have brought Anglo-French forces to Constantinople, age-old goal of Russian ambitions. Britain and France promised the straits to Russia, which consented that Constantinople should be a free port and recognized Anglo-French demands for territories in Turkish Asia. This first of the secret treaties between the Allies was never carried out because the Dardanelles campaign failed. When the British Navy finally reached Constantinople in 1918, Russia had become an enemy of the Allies. The closure of the straits in 1914 had been one of the most potent contributory causes of Russia's defeat and hence of the Bolshevik Revolution.

Italy's entry into the war. The Allies were more successful in securing the adherence of Italy. Initially Italy would have been content to remain neutral, at the price of the Austrian Trentino. A grudging Austrian promise to surrender some territories served the Italian foreign minister to extract favourable terms from the Allies (secret Treaty of London, April 26, 1915). Italy was promised southern Tirol, Istria (excluding Fiume), parts of Dalmatia and Albania, and the Dodecanese Islands (occupied since 1912), as well as a share of Turkey and of the German colonies. While Italy would have redeemed the entire Italian *irredenta,* it also would have obtained German-Austrian and Slavic populations. Italian public opinion clamoured for war on Austria, which was declared on May 23, 1915. War against Germany was delayed until August 1916.

Balkan developments. Allied negotiations with Romania, Bulgaria, and Greece proved unsuccessful. The Germans, however, persuaded the Turks to cede territory along the Maritsa River to Bulgaria and held out the lure of even more booty, especially Macedonia. Bulgaria therefore attacked Serbia (October 14, 1915), coincident with an Austro-German offensive. The Serbian Army was evacuated to the Greek island of Corfu, temporarily seized by France. Under the influence of King Constantine I, brother-in-law of Germany's William II, the Greeks played a double game and prevented Franco-British forces that had landed at Salonika from taking the offensive in support of Serbia.

Romania secured a favourable position by selling oil and grain to the Central Powers and obtaining arms from the

Allies. Having been promised Transylvania, the Banat of Temesvar, and Hungarian lands along the Tisza River (Treaty of Bucharest, August 17, 1916), Romania declared war on the Central Powers (August 27) concurrently with a strong Russian offensive against Austria. Because the Western Allies were still halted at Salonika, Romania was speedily knocked out of the war.

The Polish question. Manpower shortages, coupled with the desire to establish a belt of buffer states (*Randstaatenpolitik*) between Germany and Russia, led Germany to suggest the creation of a kingdom of Poland (November 5, 1916). This kingdom was to be ruled by a Habsburg archduke, a provision heartily disliked by the German Army. Contrary to expectation the Poles did not enlist on the side of the Central Powers. The Central Powers had stumbled into one of their greatest blunders. Since Russia would have had to provide the territory for the new kingdom, it broke off secret peace negotiations. The Allies did not fail to take advantage of the situation and authorized the establishment of a Polish national committee under their tutelage. In March 1917 the provisional government of Russia recognized the independence of Poland.

Peace attempts and the U.S. declaration of war. The stalemate of the war led to an attempt to negotiate a compromise peace through the good offices of the United States (December 1916–January 1917), but neither side was yet willing to make sacrifices and to moderate its aims. The Germans were still committed to a Siegfrieden (victorious peace) at the expense of France, Belgium, and Russia.

Francis Joseph died on November 21, 1916. His successor, Charles, recognized that disintegrating Austria-Hungary could be preserved only by reconstruction and peace. He gained a short respite because the rule of the Romanovs over Russia was cracking even faster than the Dual monarchy. The Tsar was overthrown and a provisional government took power (March 14, 1917). Democratic Russia loyally continued the war and thereby made inevitable its own destruction; the strength of the Russian Army was broken. The Western Allies might have been defeated but by the fortuitous circumstance that Germany had declared unrestricted submarine warfare and offered an anti-American alliance to Mexico and Japan. Mexico was to retrieve its lost provinces. This provocation pushed even the pacifistic United States into the war (April 6, 1917). The Entente was becoming a league of nations in which the United States rapidly acquired leadership.

Overthrow of the Tsar

Germany's military and naval hopes again were disappointed. Emperor Charles contacted Paris and promised to support French claims to Alsace-Lorraine. Because Charles slighted Italian aspirations, the Allies declined the offer. Two other Austrian attempts also proved futile. The German Reichstag voted in favour of a peace without annexations, but the army under Gen. Erich Ludendorff's influence was still chasing the Siegfrieden and established tight control over domestic and foreign affairs. Pope Benedict XV vainly appealed to the belligerents to negotiate on the basis of arbitration, reparations, reciprocal restitution, and freedom of the seas. Germany was still too strong to be peaceful.

Bolshevik Revolution. The provisional government of Russia was unable to master the chaos caused by mass desertion, famine, shortages, administrative inefficiency, and political agitation. It put down a Bolshevik uprising in the summer of 1917 but grew increasingly powerless and was overthrown by Lenin (November 7). The Bolshevik government published the secret treaties concluded among the erstwhile Allies, including a Franco-Russian pact that provided for the dismemberment of Germany. This step greatly embarrassed the Western powers. Pres. Woodrow Wilson proclaimed his Fourteen Points, which were presumed by many to supplant the secret treaties.

The Bolsheviks called for general peace on the basis of "no annexations, no indemnities," and concluded an armistice with the Central Powers (December 15, 1917). An independent (non-Communist) Ukraine was proclaimed and concluded a separate peace with Germany and Austria (February 9, 1918). Still confident of victory, the Germans wanted to impose harsh conditions on Russia. Foreign Commissar Leon Trotsky refused to sign on the dotted line, simply declared the war at an end, and called on the German Army to mutiny. The Central Powers resumed operations threatening the survival of Bolshevism and forced Lenin to accept the terms of the Treaty of Brest-Litovsk. Russia surrendered Courland, Lithuania, and Poland and granted independence to Finland, Estonia, Livonia, and the Ukraine. Kars, Ardahan, and Batum were returned to Turkey.

End of the war. The Allies succeeded in deposing King Constantine; Eleuthérios Venizélos took Greece into the war and menaced the flank of the Central Powers. The Allies had not aimed at the destruction of Austria-Hungary, but they reversed their position and recognized the Czechoslovak national committee under Tomáš G. Masaryk and Edvard Beneš. The King of Spain transmitted new Austrian peace feelers to President Wilson. In April 1918 a congress of the oppressed nationalities of the Habsburg Empire was held in Rome. Wilson declared that all Slavs must be freed from Austrian and German rule (June 28).

The unsuccessful offensives of spring and summer 1918 exhausted Germany's last military resources. A powerful U.S. army appeared in France, and the German and Austrian governments realized that the war was lost. The German general staff still held out for further resistance, but when the main German defense line was broken, concurrently with the surrender of Bulgaria, Generals Hindenburg and Ludendorff precipitated a demand for an immediate armistice (September 29, 1918).

On October 16, Charles proclaimed the federalization of the Habsburg monarchy. Wilson replied that the United States had recognized Czechoslovakia as a belligerent and had admitted "the justice of the national aspirations of the southern Slavs." Czechoslovakia, Yugoslavia, and Hungary proclaimed their independence, and Turkey ceased fighting on October 30. The Austrian Army disintegrated on the Italian front and accepted an armistice (November 3). The fate of Austria-Hungary was sealed.

Despite opposition by the U.S. general John J. Pershing, who believed that Germany was not sufficiently beaten, terms were communicated to Germany on November 8. William II abdicated and fled to The Netherlands. Hoping to obtain better terms, the army helped in establishing a German republic. Fighting ceased on November 11. The Germans evacuated Allied territory and surrendered vast quantities of war materials, including most of the German fleet and all submarines. The Allies occupied the Rhineland.

Aftermath. After British small forces had landed at Murmansk and Archangel, the Bolsheviks declared that a state of war existed between Russia and the former Allies. About 100,000 Czech troops en route to Vladivostok began hostilities against the Bolsheviks. British, French, Japanese, and U.S. troops supported the counterrevolutionary forces of Adm. Alexander Kolchak, who temporarily overthrew the Communists in eastern Russia. British forces occupied Baku and Batum; Georgia, Armenia, and Azerbaijan declared their independence. Thus, by the end of the war the Allies were in a position to destroy the Communist regime. Yet the German retreat from Russia, unthinkingly imposed by the armistice, enabled the Communists to equip themselves with abandoned German weapons and to turn the tables on the White Russians, who in the end were deserted by the war-weary and vacillating Allies.

World War I changed the face of Europe. Austria-Hungary and the Ottoman Empire disappeared. Germany was defeated, but its basic strength remained unimpaired. Russia was dismembered and had fallen into misery and chaos; but the Communists, resolute and ruthless, were gradually tightening their hold over the vast land and, in blood and hunger, were creating the basis for a supreme threat to the European West. The erstwhile unity of Europe had been struck down by the horrors of fratricidal war.

Interwar years. The peace settlement was dictated by the Allies, mainly by the leaders of the Big Three powers: Pres. Woodrow Wilson of the United States; George Clemenceau, premier of France; and David Lloyd George,

prime minister of Great Britain. It was a compromise not only between the aims of these three men but also between the avowed peace aims of the victors and the brute facts of the situation at the end of the war. Many expectations were inevitably disappointed.

The map of Europe

The total number of European states was increased. Czechoslovakia, Poland, Estonia, Latvia, Lithuania, and Finland appeared on the map; Austria and Hungary survived as two small, separate, landlocked states, while European Turkey was restricted to a narrow coastal strip along the straits. Serbia, with the addition of the southern Slavs of Croatia and Slovenia, became Yugoslavia (though it did not adopt that name until 1929). Romania was increased in size at the expense of Hungary and Russia. France recovered Alsace-Lorraine; Belgium and Poland made gains at the expense of Germany; and the area of the Saar was temporarily taken over by an international commission. Otherwise Germany remained territorially intact and set up the democratic Weimar Republic. These arrangements left large national minorities on the wrong sides of frontiers—notably minorities of Germans and Ukrainians in Poland and of Hungarians in Romania. Although the new succession states made treaties governing their treatment of such minorities, the settlement involved one-sided violations of the principle of national self-determination to the consistent disadvantage of the defeated nations.

The new states, as well as Germany, adopted democratic constitutions. Democracy was in vogue, its reputation enhanced by the victory of the Western powers, its ideals trumpeted in Allied proclamations and translated into the Covenant of the new League of Nations, which, on Wilson's insistence, was included in each peace treaty. Events proved that the unfamiliar institutions, procedures, and values of democracy were in such countries fragile and ill-fitted to survive the turbulent aftermath of war. By the end of 1926 Hungary, Italy, Portugal, Poland, and Lithuania were all under authoritarian or dictatorial regimes; and elsewhere militarist and authoritarian forces challenged the working and the survival of democratic governments. The prevalence of democracy in Europe was short-lived.

The League of Nations

The League of Nations, too, quickly disappointed hopes that it would be an effective medium for cooperation among the major powers of the world. The U.S. Senate did not ratify the peace treaties by the necessary majority of two-thirds and so did not endorse the Covenant of the League. Not only was the United States never a member of the League but the U.S.S.R. was not admitted to membership until 1934, and by then Japan and Germany had ceased to be members. The League was much weakened in power and prestige by the absence of never fewer than two of the world's greatest powers. The task of making it work devolved on the main Western Allies, France and Great Britain; and their aims and policies diverged during the testing time of the 1920s.

The social and political consequences of the Great Depression were as disruptive as the economic consequences of the war. In Europe the aftermath of the economic depression was political revolution, first and most significantly in Germany. On January 30, 1933, Adolf Hitler, leader of the National Socialist Party, became chancellor. This movement, dating from the years of turmoil immediately after World War I, had sporadically gained in strength by exploiting every difficulty of the Weimar Republic. Appealing to the aggrieved nationalism of the military and conservative classes, the embittered and ruined middle classes, and the mass of unemployed workers, it was avowedly antiparliamentary, antidemocratic, anti-Socialist, and anti-Communist. It aimed at total overthrow of the Versailles settlement and the resurgence of Germany as the dominant power in Europe. By its violently anti-Semitic racial theory, Nazism exploited, through skilled propaganda, the deepest impulses of mass hysteria and blind hatred. With this dynamic, ruthless, revolutionary movement in control of Germany, the postwar era gave place to a prewar era—a change made doubly certain by the refusal of the Western powers to perceive that this was happening.

The recovery of Germany's power in Europe began with one false move, Hitler's attempt to bring about union (Anschluss) with Austria in 1934, which was frustrated by Austrian and Italian reactions; and with one legitimate concession—the return to Germany, by plebiscite, of the territory of the Saar (March 1935). On March 16 Hitler announced that Germany would no longer consider itself bound by the military clauses of the Versailles Treaty and embarked openly on rearmament, which had already, for many years, proceeded secretly. In 1936 he similarly repudiated the Locarno agreements of 1925 and reoccupied the Rhineland with German troops.

Germany's program of planned aggressions was facilitated by other developments, both national and international. France was beset by Paris riots in February 1934, which threatened the survival of parliament, and by a series of weak governments and prolonged ministerial crises. When, in 1935, Italy attacked Ethiopia, a fellow member of the League, the Council of the League belatedly and ineffectually imposed economic sanctions on Italy (November 1935). This did not prevent Italy's conquest of Ethiopia but led in October 1936 to the Rome-Berlin Axis, agreements to cooperate made with Hitler by the Fascist government of Benito Mussolini. Meanwhile the reality of this alignment of the dictatorships was proved by their joint support for Gen. Francisco Franco in the Spanish Civil War, which had begun in July 1936. France, under the Popular Front government of Léon Blum from May 1936 onward, was preoccupied with overdue domestic reforms. It felt obliged to follow, along with Great Britain, a policy of "nonintervention," though the U.S.S.R. sent assistance to the Spanish republican forces. The U.S.S.R. itself, in these years, was engaged in far-reaching party purges and sensational trials of military leaders. In most European countries native Fascist leagues and paramilitary groups rivalled Communist militants in causing internal disorder.

Hitler's aggressions

Taking advantage of these circumstances, Hitler proceeded rapidly with his aggressions. In March 1938 he invaded and annexed Austria. Six months later he launched a campaign against Czechoslovakia, demanding cession of the so-called Sudetenland; and by the Munich Agreements of September 1938 he contrived to involve Great Britain, France, and Italy in exacting this concession from Czechoslovakia. In March 1939 he invaded the remainder of Czechoslovakia in violation of assurances given; and the following month Mussolini took Albania. In August Hitler stunned the Western powers by making a pact with the U.S.S.R., hitherto presented by Nazi propaganda as Germany's deadliest enemy. On September 1, 1939, he attacked Poland. Because Great Britain and France had meanwhile guaranteed Poland against German attack and had come to see that German aggressions could be checked only by war, they declared war on Germany on September 3. Soviet troops occupied the eastern provinces of Poland, so completing yet a further partition of that unhappy country. Italy declared itself a nonbelligerent. World War II began, therefore, as a conflict between Germany and the Western allies of France and Great Britain, backed by the Commonwealth of Nations.

World War II. There were many affinities and parallels between the two great wars that engulfed Europe within a single generation. Both began in eastern Europe; both arose from treaty obligations of great powers; both involved an initial alliance between Great Britain and France against a German-dominated central Europe; both came to implicate Germany in a war on two fronts and were won by a Grand Alliance of Great Britain, the U.S.S.R. and the United States; both changed their dimensions and their character as they proceeded, until they engaged most nations of the world; both left behind tangled problems of reconstruction and resettlement and the seeds of future conflict; of each the outcome was largely unforeseen and unintended when it began. There were also important differences. The second, more truly than the first, was a world war, for it involved prolonged fighting in the Pacific as well as the Atlantic, in Asia and Africa as well as Europe, the defeat of Japan as well as Germany (see WORLD WARS, THE).

The war in Poland ended with the surrender of Warsaw (September 27, 1939). A brief war began between

the U.S.S.R. and Finland (November 1939–March 1940), for which the U.S.S.R. was expelled from the moribund League of Nations (December 1939). The small Baltic states signed treaties of mutual assistance with the U.S.S.R. Great Britain and France, unable to influence these events, hastily prepared for Hitler's expected onslaught in the West.

Until April 1940 little happened. These early months, known as the "phony war," gave the West more time to prepare but also sapped French morale, for the strain of full mobilization proved considerable. Then with great speed German forces in April invaded Norway and Denmark and in May struck at The Netherlands and France. In Great Britain, Winston Churchill had just succeeded the prewar premier, Neville Chamberlain, unhappily associated with the Munich Agreement and the policy of "appeasement." In France, Paul Reynaud tried to rally France for the blow. By June 17 the German Army, highly motorized and enjoying good air support, had swept into France; and a new government, headed by the aged Marshal Philippe Pétain, sought an armistice. It was signed on June 22. German forces occupied northern France, Paris, and a western coastal belt. A French government at Vichy, under Pétain, remained in office to govern France under these severe restrictions.

Because Great Britain, without a foothold on the continent after the retreat from Dunkirk (May 1940), refused to sue for peace, Germany launched a concentrated air bombardment of Britain in preparation for invasion by sea. By September the attempt to gain air supremacy over Britain had failed, and Hitler in October made an abortive offer of peace. Italy, having declared war on defeated France, now attacked Greece (October 28, 1940). But Greek forces inflicted defeat on the Italians, and in April 1941 Germany declared war on both Greece and Yugoslavia. British help proved of no avail, and by May 1941 British troops were expelled from the Balkans. By now, however, movements of internal resistance had formed in all the German-occupied countries. In London Gen. Charles de Gaulle had formed a Free French movement challenging the authority of Vichy, while in Yugoslavia guerrilla war started against the Germans and Italians.

The war entered a new phase during the second half of 1941. On June 22 Germany attacked and invaded the U.S.S.R.—at first winning great victories. On December 7, Japanese aircraft attacked U.S. warships in Pearl Harbor, Hawaii. The United States declared war on Japan (December 8), and three days later Germany and Italy declared war on the United States. With dramatic suddenness the war was transformed from a war in Europe and the Atlantic to a World War involving all the great powers on all the oceans of the world. The United States, already an "arsenal of the democracies," now became an ally.

United States entry

In November 1942 the Allies, led by U.S. commanders, invaded French North Africa, and British forces won victories in Egypt. The Germans occupied the whole of France, but in the Soviet Union, at Stalingrad, they suffered their greatest defeat yet. Thereafter the tide of war began to favour the Western Allies. Italy, suffering severe defeats at sea and invaded in Sicily (July 1943), overthrew Mussolini and disbanded the Fascist Party. Marshal Pietro Badoglio took charge, surrendered on September 8, and on October 13 declared war on Germany, which occupied northern Italy.

The Western air forces, with strong bases in North Africa and southern Italy as well as in Britain, launched concentrated bombing attacks on Germany, in preparation for invasion of German-occupied Europe. On June 6, 1944, the Allies landed in Normandy and by August 25 liberated Paris. At the same time Soviet forces, having expelled the Germans from Russian soil, reached the suburbs of Warsaw. The final year of the war brought a mighty convergence of Allied armies upon Germany—Soviet from the east, Anglo-U.S. from the south and west. The Soviet Army occupied Romania in August 1944 and Bulgaria (although that country had remained neutral) in September. Greece lapsed into civil war between Communists and anti-Communists. As a result of the Yalta Conference (February 1945) of Churchill, Pres. Franklin D. Roosevelt, and Stalin, Poland fell virtually under Soviet control, yielding eastern territories to the U.S.S.R. but receiving Silesia, Pomerania, and the bulk of East Prussia from Germany. In Yugoslavia Tito, the Communist leader of resistance against Germany, was acknowledged ruler of the country. Stalin installed a Communist government in Hungary soon after Yalta. Soviet power had replaced German power in eastern Europe even before the war ended.

On May 2, 1945, fighting stopped in Italy; on May 7 the German high command surrendered unconditionally. Hitler had committed suicide when the Russians reached Berlin. The extensive campaigns against Japan in the Far East and the Pacific continued. Two days after the first U.S. atomic bomb had been dropped on Hiroshima (August 6) the U.S.S.R. declared war on Japan and invaded Manchuria. Japan surrendered on August 14. The settlement had to be global in character as war itself had been. To achieve such an aim a new international organization, the United Nations, was prepared at the conference of San Francisco (June 1945). Its charter was signed by representatives of 50 nation-states. Special agencies, such as the UN Relief and Rehabilitation Administration (UNRRA) of 1943, already existed to deal with immediate relief work.

Communism and Europe. The wartime agreements of the Allies envisaged a lengthy postwar period of concerted action for a wide variety of tasks: during the "temporary period of instability in liberated Europe" (Yalta, February 1945); for the joint administration of occupied Germany (Potsdam, August 1945); for conclusion of treaties of peace with other former enemy states; and, within the United Nations and its several agencies, for continued cooperation of many kinds. This policy failed after two years, but during those years it set the political pattern of the postwar world.

It settled, first, the pattern of state frontiers. During 1946 the major powers prepared peace treaties with Finland, Hungary, Romania, Bulgaria, and Italy; these were signed in February 1947 in Paris. The provisional settlement with Germany was dictated at Potsdam. The general outcome was that the U.S.S.R. acquired direct control over the whole belt of the eastern marchlands comprising Petsamo and eastern Karelia; the three Baltic states of Estonia, Latvia, and Lithuania; the northern part of East Prussia; the eastern part of Poland; the Carpatho-Ukrainian area of Czechoslovakia, Bessarabia, and northern Bukovina. Apart from adjustments to the Polish and Czechoslovak and the Italian and Yugoslav frontiers, prewar frontiers were broadly restored. In western Europe the Saar was at first (1947) established as an autonomous territory in economic union with France, but in 1957 it was reunited with West Germany. Germany, at first divided into four separate zones of occupation by Soviet, U.S., British, and French forces, was in 1949 constituted as two distinct states, the Federal Republic of Germany comprising the three western zones, and the German Democratic Republic comprising the eastern zone. Berlin, itself divided at first into four zones and jointly administered by the four powers, was similarly split into two zones—West and East—by 1949.

A divided Germany

The policy determined, secondly, the pattern of postwar regimes in Europe. In most liberated or conquered countries governments were at first coalitions of Communists, Socialists, and Christian Democrats or Peasant parties. In most Western countries (including France) and some Eastern countries (Hungary, Czechoslovakia, Poland) such governments were set up by more or less free elections. They usually carried through important reforms, redistributing the land or nationalizing industries and instituting provisions for social welfare. In these years the welfare state became the prevalent pattern in Europe. In Great Britain active Labour governments (1945–51) achieved comparable reorganization.

After 1947 the contrasts between East and West became sharper in domestic as well as in international affairs. In Eastern lands, under Soviet control or influence, non-Communists were removed from power and the whole area east of the Elbe and the Adriatic, with the exception of Greece and Turkey, fell under the control of Communist parties. In the West the contrary process occurred.

Communist parties were excluded from a share in power in France, Belgium, and Italy in the spring of 1947 and in the Federal Republic of Germany in 1949. They had never been strong in the Scandinavian countries and did not appear in Portugal and Spain until the 1970s. Internal economic policies changed correspondingly in the two halves of the continent: increasingly restorative and moderate in the West, increasingly revolutionary in the East. In Germany economic divergence culminated in the introduction of separate currencies in June 1948. By 1949 the era of Cold War between East and West had set in, and the iron curtain from Stettin to Trieste divided Europe into two worlds.

North Atlantic Treaty Organization

In April 1949, the North Atlantic Treaty Organization (NATO), a defensive alliance providing for common strategy and joint military forces under a unified command, was set up. From the first it included the United States, Canada, the United Kingdom, France, Belgium, Luxembourg, Denmark, Norway, Iceland, Italy, The Netherlands, and Portugal; later it was joined by Greece and Turkey (1952) and the Federal Republic of Germany (1955). The Warsaw Treaty Organization (May 1955) created its formal counterpart for eastern Europe, comprising the Soviet Union, Poland, Czechoslovakia, Hungary, Romania, Bulgaria, Albania (which withdrew in the 1960s), and the German Democratic Republic. The settlement was one of partition and armed deadlock.

The partition of Europe, thus consolidated by midcentury, brought about new relationships between Europe as a whole and the continents of Asia and Africa. The colonial revolution, dramatically advanced by the war, had already gone far. Former colonial territories asserted their independence and joined the United Nations as separate states. The diminished status of Europe in the world was emphasized by the growing importance in world affairs of the new Afro-Asian states.

Awareness both of the threat of Communism and of this changing balance of world forces fostered, during the 1950s, movements for closer European union. The Council of Europe dated from 1949, the European Coal and Steel Community from its signing in 1951. Western European Union (WEU) was set up in 1955, organizations for a Common Market, or European Economic Community (EEC), and for joint development of atomic energy (Euratom) followed in 1958. In economic, military, and social cooperation the peoples of western Europe equipped themselves with abundant machinery and facilities. By the early 1960s their economic recovery seemed complete, and productivity surpassed all former levels. Social services and welfare provisions made for a higher standard of living and of social security. Despite lavish expenditure on armaments and national defense, the peoples of western Europe experienced considerable material prosperity.

Throughout the late 20th century the relative ascendancy over Europe of the United States and the Soviet Union was emphasized by such developments as the space race between them (in which Europeans could scarcely afford to join), the thrust of U.S. economic expansion into Europe, the invasion of Czechoslovakia by Soviet and other Warsaw Pact troops in 1968, and the repressive measures in Poland in 1981–82. Clearly, Europe no longer counted for as much in the world as it had before 1939. Yet this ascendancy, too, was in fact counterbalanced by other world forces: the rise of Communist China as an enemy of both the superpowers, severe internal tensions within the Communist world and within the United States itself, and the readiness of various countries to exploit such conditions for national advantage. The world balance of power was shifting; it was not destroyed. (D.Tn./Ed.)

MODERN ECONOMIC GROWTH AND DEVELOPMENT

The course of industrialization, 1870–1914. *The changing balance of economic power.* The last part of the 19th century saw the further ramification of the Industrial Revolution into the peripheral European lands, the offshoots of Europe overseas, and—for the first time—a non-Western society, Japan. The result of this geographical extension, combined with differing national rates of productivity change and industrial growth, was a drastic revision of the international balance of economic and political power.

Within Europe, these decades brought an important realignment of economic power as the first follower nations (Germany, France, Belgium) completed their industrial revolutions and new entrants (Sweden, Italy, Holland, Hungary) made their appearance. Britain, the first industrial nation, continued to develop and grow, though more slowly than its rivals. By the 1890s Germany had passed Britain in iron and steel manufacture, and, from then on, the gap between the two countries widened steadily. The result was a shift in the balance of political power that generated international anxiety and conflict and contributed to the outbreak of World War I.

The pressure of international competition was felt most keenly in Great Britain, which, as the dominant industrial power, had most to lose. The sense of loss was the stronger because these were decades of slow growth of manufactures (1.8 percent a year), declining prices (the wholesale index for manufactures fell from 127 in 1873 to 71 in 1894), protracted slumps, and small profits—hence, the traditional designation "the great depression."

Business cycles

Some scholars have viewed the great depression as a European phenomenon, but the course of economic activity in the continental countries was quite different from what it was in Britain. German industry did suffer a severe setback after the crash of 1873, a typical catharsis after some three years of enthusiastic company promotion (the *Gründerjahre*) fuelled by the large French war indemnity. Beginning in the 1880s, however, rapid growth resumed, and profit levels picked up, with no further serious interruptions until the crisis of 1900–02. France, by way of contrast, had much less trouble in the 1870s, was hard hit by a financial crisis in 1882, had another, smaller shock in the mid-1890s, and then another, more protracted, in the first years of the new century. On the whole, these were years of below-average growth for manufactures.

A second Industrial Revolution. Insofar as the Industrial Revolution spread during these decades to newly industrializing countries, its outlines were substantially the same as they had been earlier; that is, the first advances typically took place in food processing and textiles (cotton especially), with further extension into metallurgy, machine building, and chemicals. Within each of these branches, the techniques and equipment introduced were largely the classic elements: spinning machines and power looms, coke blast furnaces, and rolling mills. But the equipment itself had grown bigger, faster, more accurate and dependable, while the array of techniques had vastly proliferated, so that newly industrializing nations adopted, alongside these classic items, such newer ones as the steam hammer (an invention of the 1840s), the sewing machine, the milling cutter (for metal as well as wood), and the new technology of steelmaking. As a result of these changes, the threshold scale of production had grown considerably, and each piece of equipment cost substantially more. In short, though the lineaments of these later industrial revolutions were the same as earlier, the content was quite different.

In the meantime, the technology of the advanced nations continued to evolve in the same direction, although the gains achieved in the older branches were shrinking relative to previous advances. The spindles could turn a little faster; the furnaces could be built higher; but the increment entailed a smaller proportional increase in profit than the radical changes of an earlier era. The place of the old staple industries was shrinking, and the rate of growth of a given economy depended on its ability to develop newer lines of production.

These constituted, toward the end of the 19th century, an innovational cluster important enough to warrant perhaps the designation of a second Industrial Revolution. As before, the critical advances occurred in the fundamental sphere of energy conversion and distribution: on the one hand, the production and use of electricity as a source of motive power and, on the other, the invention of the internal-combustion engine as a prime mover. The first had revolutionary consequences for both industrial technique and domestic life. Not only was electrical transmission of energy more efficient and safer than conventional direct

Electrical power

drive by shafts and belts but it also could be effected over long distances, so that the production of energy could be centralized and power distributed to small units (households and shops, as well as mills and forges) as needed. Electrical current, moreover, was far more versatile than steam power, for it could be used to produce not only motion but also light, sound, and heat, and it made possible a revolution in communication (telegraph, telephone, radio, television, radar, telex, etc.), as well as some notable advances in chemical manufacture. Whereas steam power was necessarily concentrated in large production units, hence, in a few power-intensive industries, electric motors were useful and profitable in a much wider range of activities; thus, in 1889, the five leading power-using branches (lumber, food, iron and steel, textiles, paper) held 73.5 percent of total primary-power capacity in manufacturing; by 1929 this share had fallen to 58 percent. Germany took the lead both in production and use of electrical power and in the production of electrical equipment. German industry was marked by thorough exploitation of by-product energy produced by heat-generating industries, large distribution nets with high-use factors, and early standardization of current characteristics.

Internal-combustion engine

The second major innovation in the generation of power was the internal-combustion motor, so called because the driving force is obtained by controlled explosions within the motor itself. The first such device was Étienne Lenoir's gas engine of 1859, but the critical conceptual advance came with Alphonse Beau de Rochas' four-stroke cycle, which found application in Nikolaus August Otto's "silent" gas engine of 1876. This was a tremendous success from the outset, for it was far more efficient than the steam engine when working intermittently or at less than full load, conditions typical of small industry. The gas engine was tied to its source of fuel, but the internal-combustion motor was compact and had potential mobility. These possibilities were realized with the adoption of oil fuel. The result was a revolution in transportation (marine engines, the airplane, and, above all, the automobile), which liberated land transport from rail routes and the individual traveller from dependence on common carriers. The impact of the automobile was only beginning to be felt in the period before World War I; but the introduction of the cheap passenger car (Henry Ford's Model T, from 1908) was a promise (or warning) of things to come. The automobile quickly became the most important single product of the manufacturing industry, generating investment and growth by its demand for raw materials and highways (including bridges and tunnels) to match the burgeoning volume of traffic. Output of petroleum products, rubber, cement, asphalt, steel, glass, and even textiles waxed (and sometimes waned) in sympathy with the demand for and use of automotive vehicles of all sorts, to say nothing of a large service sector (insurance, law and litigation, medical care, auto repair, garages, and parking lots) devoted partially or exclusively to meeting the needs and correcting the misdeeds of auto users. At the same time, the process of manufacturing automobiles called for innovations in materials and techniques important in themselves and susceptible of application in other industries: line assembly, new alloys, special-purpose tools, advances in grinding and stamping, and standardization of parts.

Importance of the automobile

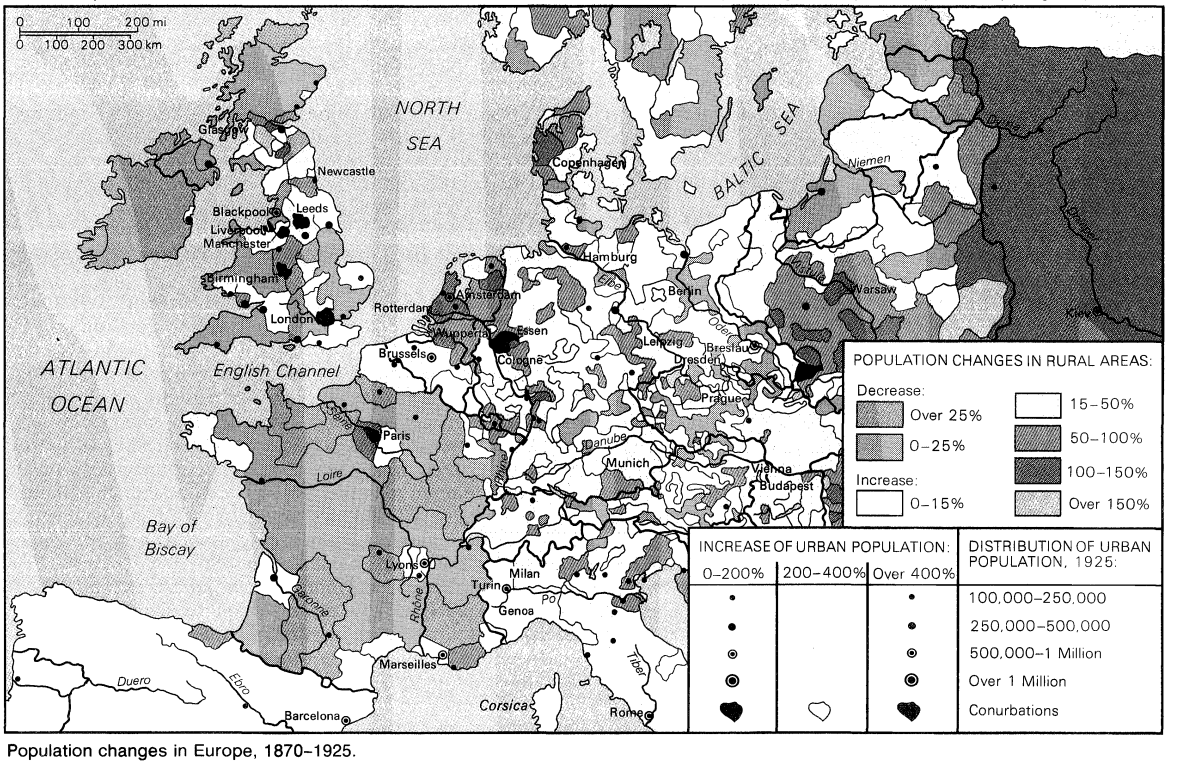
As important as the automobile industry came to be, two other new branches of manufacture can be compared to it in range of influence and potential for growth. One was organic chemicals and the manufacture of synthetics. Again, the pre-World War I years gave just a hint of what was to come: celluloid (John Wesley Hyatt, 1868); artificial silk or rayon (Hilaire, comte de Chardonnet, 1889–90); bakelite, the first synthetic resin (Leo Hendrik Baekeland, 1909); cellophane (J.E. Brandenberger, 1912). These artificial fibres and plastics were the first fruits of what would be a new kind of industry, based on increasingly heavy investment in laboratory research and turning out new products to order, primarily for consumer use.

The second was the equally versatile electronics industry, in its infancy in the early years of the 20th century, when its only application was still wireless communication. The key innovation here was the tube or valve, so called because it channelled the flow of electric current in one direction and made possible the construction of complex circuits. Synthetics and electronics did not fully develop until the interwar years and the post-World War II period and might better be seen as part of a new major cluster of innovations (including atomic energy and computers) that some regard as constituting a third Industrial Revolution.

Along with these and other innovations in technique and product went new patterns of industrial and business organization. The old trend to bigness continued as better but costlier production equipment raised the price of entry

The trend toward bigness

Adapted from H.C. Darby and Harold Fullard (eds.), *The New Cambridge Modern History Atlas*, vol. 14 (© 1970), Cambridge University Press; maps © 1970 by George Philip & Son, Ltd.



Population changes in Europe, 1870–1925.

and offered new economies of scale, while improvements in communication enhanced the possibilities of centralized management. Economies of scale—in marketing and administration as well as production—were only part of the story, however. Equally important were market considerations, in particular, the desire to stabilize profits and minimize competition, either by absorbing or eliminating competitors or by organizing cartels, trusts, and similar combinations in restraint of trade. The merger and cartel movements were especially strong in Germany. By 1905 there were some 366 industrial cartels operating in Germany; some of them controlled prices, others output, many both. In Germany, the trend to bigness and combination was so powerful and the role of the so-called great banks (themselves the result of numerous mergers) so influential that contemporary observers thought they saw the appearance of a new stage in economic development, the era of finance capitalism.

Finance capitalism

The term was somewhat premature. Although similar trends were to be found in the other major European industrial economies, they were substantially weaker. In both Britain and France, for example, the stress on family ownership and identification posed a barrier both to financial intervention in industrial management and to disciplined participation in cartel agreements. In Britain, when mergers did come, they were often a response to contraction rather than an act of expansion: independent firms reluctantly combined their resources to meet falling prices and profits and then spent years fighting over the appropriate allocation of authority and returns.

Industrialization on the eve of World War I. The position of the different nations in this process of social as well as economic transformation may be measured by the position of industry relative to agriculture and the ratio of urban to rural population. In regard to the first, one can discriminate among four categories.

First, there were countries in which the industrial labour force was already greater than that in agriculture: Great Britain, which had reached this position by 1820, is the best example, but Belgium had pushed its industrialization almost as far. The United States and Germany, in spite of substantial advantages in agriculture, had long since entered this group. France had not quite made it and would not do so until after World War II. Each of these countries had increased product per head by three to five times over the course of the preceding century.

Second, there were countries whose agricultural population was still about twice as great as its industrial: Sweden, Italy, and Austria all fell into this category. Each nation had developed by this time a large, modern industrial sector, and Sweden and the Czech lands, particularly, had enterprises that were technologically at a level with the best European practice.

Third, there were countries that had entered on the Industrial Revolution but were still primarily pre-industrial. Russia was a good case in point, with a few completely modern plants, especially in the young iron-and-steel industry of the Donets Basin. Hungary also had, in the Ganz electrical works of Budapest, one of the more progressive enterprises in Europe. But these were predominantly peasant societies; skilled industrial labour was scarce; and industry was still dependent on strong tariff protection and artificial support.

Fourth, there were countries almost exclusively dependent on agriculture: such Balkan nations as Greece and Bulgaria, the colonial lands of Asia and Africa, and the more backward countries of Latin America—in short, what would later be called the Third World. With rare exceptions, industrial production consisted almost entirely of domestic handicrafts and artisanal work.

Concomitant with the shift of resources to industry went a movement of population to urban centres. The development of a predominantly urban society entailed substantial costs to physical and psychic health, while the convenience of concentration was offset by the impediments of high density of living and movement. Even so, urbanization contributed to the increase in productivity and wealth by intensifying contacts and information flows between innovators in knowledge and its applications and by nourishing and propagating new attitudes toward family size and population control. (D.S.La./Ed.)

The aftermath of World War I. The outbreak of war in 1914 ended a distinct era in economic history and ushered in a period of instability, uncertainty, and unexampled change that could hardly have been foreseen before the conflict began. Trends that might have been gradual were telescoped into a few years. A further impetus was given to the concentration of production in large-scale units and to the decline of competition. State intervention in the economy, prominent for the first time under modern conditions during the war, was to assume a more definite and permanent character. Technological change became more firmly linked to the giant advances in the physical sciences and permanently molded life in the advanced regions of the world.

Only the most technologically and industrially advanced nations could wage wars of the type of 1914–18 and 1939–45. Germany's ability to hold out so long was a tribute to its industrial strength as much as to its military competence in both wars. The decisive role of the United States as an arsenal and its rapid industrial growth in both wars is again highly significant. France held out in World War I thanks to its allies' economic support; its collapse in World War II was as much economic as it was military. With or without the atomic bombs, Japan would have been defeated because of its overstretched resources.

In this period, economic history cannot be separated from military history, and into both the major wars there entered deep economic rivalries. War also provided an overwhelming stimulus to maximize production and make full use of scientific knowledge. Mechanization, rationalization, and standardization were imposed by the conditions of war production, and the interruption of normal trade channels intensified the search for substitute materials and the development of synthetics, from which are descended some of the fast-growing industries of the period. The "new industries" early in the century expanded rapidly.

The conditions of war production

The diminution of European power. It was the world economy that experienced the most profound changes after the war. The most outstanding was the financial and industrial upsurge of the United States and the decline of Europe, transformed from a creditor to a debtor. The world monetary and trading system centred on Europe had been seriously dislocated, and frontier changes had increased the tariff barriers dividing the Continent. The Bolshevik Revolution had taken Russia out of world capitalism and threatened to spread westward in the postwar atmosphere of discontent and near chaos. Finally, industrialization had gone ahead in other parts of the world, closing old markets and rearing up new competitors. Most ominous of all was the emergence of Japan as a vigorous competitor in world markets.

Of immediate significance in the postwar decade was the need to restore the old pattern of trade and payments. This was taken to mean, first, the restoration of the gold standard upon which prewar currency stability and convertibility had been based. But war finance had caused inflation in the belligerent countries, and there could be no return to gold without currency stability. Second, Europe was unable to resume its prewar position as an international creditor. U.S. capital played a vital part in European reconstruction and subscribed heavily to the international loans that enabled Germany to make scaled-down reparations payments. Stabilization and expansion in the 1920s was thus underpinned by an outflow of funds from the United States. Funds returned to take advantage of the capital gains realizable in the U.S., and when the crash came in 1929 one of the main props of the central European banking structure was removed, and the stage was set for the financial collapse of 1931.

The role of U.S. capital

War finance had been inherently inflationary, and, once controls were removed, prices began to soar. Moreover, social breakdown in eastern Europe and the Soviet Union had resulted in money printing and inflation. The occupation of the Ruhr by French troops in 1923 precipitated a hyperinflation in Germany that broke all records. In the final stages Germany resorted to the printing press on a massive scale, and the rise in prices was such that workers

collected their wages in suitcases and spent them at once. In November 1923 a determined and successful effort was made to stabilize the currency, and Germany entered a new period. Industry had meanwhile wiped off its debts and made considerable windfall profits. Thanks to this and to the inflow of U.S. funds, industry was modernized and expanded. Inflation had disastrous effects for those who held their savings in state bonds or fixed interest securities or who lived on fixed incomes. At the same time, speculators and profiteers did well, and some made fortunes overnight. The "new rich" of the inflationary years prospered; their social juxtaposition with impoverished middle-class people helps to explain the political upheavals that followed. The outstanding feature of the inflation was its undermining of the old virtues of thrift and savings, which had been capitalism's most valuable moral supports.

The changing world picture. The primary producing countries had boomed during and immediately after the war and had increased their capacity. The European recession that began in 1921 after the postwar restocking boom was a sign that things had changed. Most of the producing countries were in a colonial or dependent relationship to the advanced European countries, and in a number of them a nationalist movement had appeared. Where some industrial growth had taken place, as in India, protection of the home market and government encouragement of industrialization were needed.

Nationalist movements

Some economists have linked Europe's interwar difficulties with the disappointing performance of primary producers. Had the European countries adapted to changes in technology and world demand, they would have increased their demand for primary products and enabled their producers to buy more manufactured goods. As it was, world trade was the worst victim of the war and the biggest casualty in the Depression of the 1930s. To make matters worse, the industrial countries protected their own producers, and, when the slump came, there was a general retreat to the home market with higher tariff barriers, quotas, and administrative devices to keep imports down. Primary producers responded by encouraging their own industries as well as keeping up prices by restrictive schemes.

The beginnings of social welfare. The economic experience of the war years did not shake the general assumption that, in peacetime conditions, state intervention was unnecessary and harmful. But it had been shown that in certain circumstances business might benefit from the activities of the state. Until the Depression of the 1930s, however, there was no desire to see the state permanently endowed with more extensive economic powers. It was not the success of state intervention in the war but the breakdown of market forces in the 1930s that opened the way for regulated capitalism.

The extent of state intervention varied greatly; some countries had operated the railway system for many years, and there were state monopolies for fiscal purposes. But the main economic intervention continued to be the tariff. Even Britain had imposed protective tariffs on luxury items during World War I and in the 1920s began "safeguarding" industries from unfair foreign competition. There was also a steady underlying trend toward greater state involvement in the economy, influenced by the growing Socialist and trade union movements. The way was being prepared for the wider intervention by the state characteristic of the Depression decade.

The use of tariffs

State provision for social welfare had a much wider measure of acceptance in the 1920s and steadily became more systematic. It was understood that the sanctity of private property and the pursuit of profit might be threatened if society did not make adequate provision for its casualties. By the 1920s most states had social security or insurance systems of various degrees of comprehensiveness. Increasingly, such welfare provision became a matter of legislation laying down minimum standards. Wartime experience, in which all members of a society were taken to be in solidarity against a common enemy and thus entitled to its support, exercised an important influence on the development of social security systems.

The state's tendency to assume social responsibilities was persistent and continuous in the period after World War I, and, with the establishment of the International Labour Office in Geneva, a beginning was made in establishing uniform standards between countries. Partly under labour union pressure, industrial regulation became more comprehensive and better enforced. Public and local authorities were also assuming greater responsibility in transport and the provision of water, gas, and electricity; in education; and in housing, made necessary by the chronic and widespread postwar shortage. Investment in human capital, if not yet a common phrase, was being taken more seriously than ever before; but, like all similar trends, it was uneven.

During the 1920s, as recovery and inflation took place, production and productivity again began to grow in the advanced countries, and incomes rose. Meanwhile, the new technologies began to influence social life more fully. Road transport made its effects felt through the bus and the truck; rural areas were brought into closer contact with urban influences; and personal and labour mobility increased. Marketing and distribution changed as farm produce could be collected more regularly and moved more quickly to urban markets. Electric power brought new flexibility to industry and opened an industrial vocation to hitherto predominantly agrarian areas. The cinema and the radio transformed entertainment patterns as well as the supply of information or propaganda. Professional sports became a growing industry. Service trades of all kinds grew most rapidly in the advanced countries with high incomes.

The new technologies

This expansion of the "tertiary sector" was one of the most notable trends of this period. Consumers, with increased incomes, spent more on services. Former household activities became specialized marketed services; retail outlets multiplied; old-style services expanded, and new ones appeared. As the activities of governments became more complex, so their employees increased; and even business employed a growing bureaucracy of its own. In other words, the white collar element in the labour market grew and was symbolic of changes in life-style in societies with a relatively high and growing per capita income. Expectations grew, too, even faster than possibilities for their realization. Workers became better organized and more demanding, installment selling enabled people to spend ahead of their incomes, and the Depression caught all these developments in full flood.

The Great Depression. The stabilization of capitalism after World War I was short-lived, for the world economy plunged into the worst depression in its history. Although heralded by the Wall Street crash in the autumn of 1929, it was some years before it became apparent that this was no ordinary cyclical downturn. It afflicted every capitalist country, and its depth and length still challenge those seeking its explanation. Even on the outbreak of World War II, despite the recovery from the trough of 1932–34, the Depression had not been conquered. So far from new government policies for full employment and economic growth being effective, it seemed rather that purely national remedies, the resort to beggar-my-neighbour policies and, more especially, the drive to rearm—which arose out of the political upheavals of the Depression—prepared the way for the new war.

What does seem to be accepted is that, although depression was centred in the United States, it was the special and intimate relationships established between the U.S. and other capitalist economies after World War I that accounted for its rapid diffusion and harsh impact. It is equally apparent that the stabilization that had taken place in the other countries and the boom conditions of many, in the late 1920s, concealed serious weaknesses. The British economy's failure to resume capital exports in the old way, the decline of its basic industries, the consequences of the overvaluation of sterling—all weakened the world economy and thrust greater responsibilities upon the United States. The investment boom in European countries following reconstruction could only continue so long as markets could be found; a cyclical downturn was inevitable sooner or later. The problems of the primary producers (now weak sellers as a result of overproduction)

International economic weaknesses

reduced the market for manufactured goods and slowed their growth rate.

In the return to gold, some currencies had been overvalued, others undervalued, and the Depression came before adjustments could be made. Moreover, international financial relations were no longer on prewar lines. War debts and reparations weakened capitalist countries, leaving the United States as the major creditor and financier of the 1920s' boom in Europe. Once the U.S. economy slumped, there was no prospect of maintaining prosperity elsewhere.

The collapse of the world market. The contraction of world trade was the most general expression of the crisis, hitting hardest the primary producers and those industries depending upon exports. Thus, outside the United States, the Depression was worst in Germany and Britain. The sharpest contraction was in the export trades and the capital goods industries, as well as in shipping and shipbuilding. Activities turned mainly toward the domestic market suffered less, and so "sheltered" and "unsheltered" industries were distinguished. The former, least affected by depression, were also the easiest to help. The recovery of the export and capital goods industries was much more difficult to engineer.

For most countries, the more prosperous conditions had been before 1929 the sharper was the decline in the following years, but a large agrarian sector—as in France—could provide some shelter from the blizzard. Germany was the worst hit European country, and there the Depression was a social and political calamity. Britain was aided by the abandonment of gold and the devaluation of sterling; the adoption of protection that stimulated home industry; a delayed development of the technologically more advanced industries; a building boom assisted by cheap money; and the ability to take advantage of the low prices of imported foodstuffs and raw materials while using accumulated reserves to avoid a balance of payments problem.

The Depression was traumatic because it contradicted deeply held beliefs in unlimited expansion and also brought poverty, hardship, and hopelessness to millions. The Depression had such severe social consequences that even reluctant governments, formerly tied to laissez-faire ideas, could not stand aside. The price of nonintervention was social revolution, and intervention was, or could be, a weapon of counterrevolution, as under the Nazi regime in Germany. Intervention, therefore, took place pragmatically, and its forms were at first "orthodox" in that they consisted of old, well-tried methods. First, the tendency toward economic nationalism was reinforced as governments imposed tariffs or raised existing ones and added quotas and other devices to protect home producers. As the Depression continued, protection tended to be used increasingly aggressively to bargain with other countries and penetrate foreign markets. The disintegration of multilateralism, which was often replaced by bilateral trading agreements, accompanied the decline of world trade. The immediate effect of these measures was to reduce the volume of international trade and create trade war.

Parallel to this went the breakdown of the gold standard. In the autumn of 1931 the overvalued pound sterling was driven off gold, starting a succession of devaluations as countries learned that an undervalued currency gave a temporary advantage in the struggle for markets. Currencies became subject to government action, and some form of exchange control had to be established if balance of payments difficulties were not to be exaggerated by speculation. Countries like France, which delayed establishing exchange control, did so to their cost. Nazi Germany, on the other hand, by means of rigid controls, could operate parallel exchange rates to its advantage and at the expense of weaker trading partners.

The conditions of recovery. The Great Depression of the 1930s shook capitalism to its foundations and was the major formative influence on political and social as well as economic life. Already undermined by World War I and its aftermath, many of the old certainties derived from 19th-century liberalism disappeared. Despite millions of unemployed throughout the decade—in many of the most advanced areas of the world—and the triumph of Fascism in Germany and the subsequent drive to war, most years of the 1930s were, however, years of low-level recovery. Indeed, the slump itself had prepared the conditions for recovery through the normal cyclical process: after a few years, in which investment falls to a low level or even becomes negative, stocks are used up, equipment grows obsolete and has to be replaced, and prices reach bottom, the possibilities for profitable investment begin to improve. Unsound firms have disappeared, the financial excesses of the boom have been liquidated, a shift of resources from the unprofitable sectors has been affected, old plant has to be renewed, and confidence revives. Thus, the classic conditions for recovery began to operate independently of government action.

Recovery came slowly and in many countries was far from complete by World War II. Until there was a revival of world trade, the primary producers could not raise their incomes. On the other hand, in the advanced countries, even those as export dependent as Britain and Germany, it was possible to find a substitute for exports. In Germany this was chiefly war production. In Britain, although rearmament in the late 1930s helped recovery, it had earlier depended upon an expansion of the home market, the shift to new industries, and the continuous rise in service activities, even in the "depressed areas."

As a major importer of foodstuffs and raw materials, Britain was able to take advantage of the deterioration in the terms of trade of primary products. This enabled consumers to spend more of their income on nonfood items and expanded the consumer goods and services market, creating new opportunities for profitable investment and employment. It assisted manufacturers directly and consumers indirectly through the relative cheapening of imported raw materials. Given that this took place with a resumption of technological advance and innovation, Britain could enjoy a limited recovery. Almost all persons could thus improve their real income, at the expense, in part, of primary producers abroad. Where those primary producers were to a large extent at home, as in France, these conditions for recovery did not exist.

All governments realized that something had to be done to relieve what amounted to a national catastrophe. The Soviet Union's example had begun to turn thoughts toward planning, even of capitalism, as a possibility. In the early years of the Depression the instinctive reaction of politicians was still strongly orthodox. In Britain, proposals for positive steps to overcome unemployment through public works were rejected by a Labour government. The watchword virtually everywhere was to keep the budget balanced and avoid interference with free market forces. Policy was thus generally deflationary: government expenditures were vigorously reduced, and the more consistently such remedies were applied, as in France, the more they aggravated the disease. In some cases, however, governments did inadvertently run budget deficits through emergency action, and where they did so they began, possibly unwittingly, to assist recovery.

The orthodoxy of political reaction

The industrialization of the Soviet Union. The Bolshevik Revolution, which in 1917 took Russia out of the capitalist world, and Russia's transformation into a leading industrial power have had enormous significance in the 20th century. In the early years the very existence of the new regime weakened capitalism; and later the viability of the alternative system was demonstrated in the carrying out, through nationalized property relations, of planned large-scale industrialization. Whether or not Premier Joseph Stalin's tyrannical methods were necessary to the process, the results were impressive, indicating to less developed countries that there was an alternative to capitalist industrialization. Elsewhere, too, planning was turned from vague concept to practical reality, acquiring exceptional prestige even in capitalist countries.

The early days. The economic steps taken by the Bolsheviks in their first year included nationalization of all industries employing more than five people, of banks and foreign trade, and of the land. A large part of the country, however, was outside the new Soviet government's control, and civil war continued until 1920. There was,

in fact, virtual economic breakdown, with many factories closed or producing only a trickle. Priority was given to the front for supplies available in what was substantially a siege economy named "War Communism." The government appropriated what revenue it required either by requisition or by printing money. The black market engulfed supplies, stocks were hoarded, and galloping inflation began. As production fell and goods disappeared from the shops, workers who had not been drafted into the army or employed in the new administration drifted back to the villages. The hardships of the period were only made bearable by promises of a better life in the future. Discontent, already growing, now became more open. The mutiny at the Kronstadt naval base in February 1921 was the final danger signal. War Communism was abandoned and replaced by the New Economic Policy (NEP).

The NEP was generally admitted as a retreat. Its basis was that it permitted private enterprise in internal trade and to some extent in industry. It also encouraged the peasants to produce a surplus for sale and offered opportunities for their enrichment. Nepmen (private businessmen) and kulaks accumulated capital.

Central planning. It was axiomatic that the U.S.S.R. could only overcome its historical backwardness through industrialization and that this should take place under the control of a central planning agency and on the foundation of a nationalized industry. The NEP stood by this principle.

The establishment of Gosplan

In fact, the State Planning Committee (Gosplan) was set up in February 1921 to "work out a single general state economic plan and methods and means of implementing it." What was at issue was the rate of industrialization, the tempo of growth, which could not in practice be separated from the path that was to be chosen.

After Lenin's death in 1924, Joseph Stalin strengthened his hold on party and state apparatus and formulated the theory of "Socialism in one country"—concentration on building up the Soviet economy at the expense of the international goals of Communism. He was supported by Nikolay Ivanovich Bukharin and others on the right wing of the party, which assumed that NEP would last for a considerable time. Industrialization, therefore, could take place only by slow stages in which the peasantry was encouraged to market its surplus and nationalized industry produced the consumer goods it required in exchange. These perspectives were opposed by the left opposition, whose leader was Leon Trotsky. "Socialism in one country" was rejected as a revision of Marxism because it placed internal interests before those of world revolution. The policy of concessions to the kulaks and the adoption of a slow industrialization tempo were equally condemned, and a program for more rapid industrialization and planning and the encouragement of voluntary collectivization was substituted. Finally, the left opposition was defeated, and Stalin emerged as the master of the party apparatus. Free debate was stifled, and the dead hand of orthodoxy gripped Soviet political economy.

Meanwhile, the New Economic Policy enabled recovery to proceed fastest of all in agriculture but also in industry. Once the prewar levels had been attained (from about 1925), decisions had to be made on the rate of growth. This depended upon how much grain surplus the peasants were prepared to part with. And that in turn depended upon the ability of industry to supply industrial goods. The main emphasis would have to be upon consumer goods production, and capital formation would be a dependent variable and not the main determinant. If, however, economic policy continued along these lines, industrialization would be slow, and the U.S.S.R. would be vulnerable to capitalist pressure or even attack. Meanwhile, the growing prosperity of the kulaks and the Nepmen would undermine the nationalized property relations upon which the state was founded.

Stalin's change of course

These dangers prompted Stalin to make a brusque alteration of course, taking over much of the policy of the defeated left opposition and applying it with a brutality they had never advocated. From 1928, therefore, the Soviet Union suffered a new period of strain and sacrifice in which, under conditions approaching civil war, agriculture was collectivized and the first stages of planned industrialization were forced through. From this caldron the U.S.S.R. emerged to face the German attack of 1941 with a much more powerful industry than before. The basis had been laid for it to become the second largest industrial power, but the social and human price had been tremendous.

The drafting of the First Five-Year Plan was begun by Gosplan in 1927; the starting date was from October 1 of the following year, but the campaign did not get under way until 1929. At the end of that year, the plan targets were revised upward at about the same time that Stalin ordered forced collectivization.

A whirlwind descended on the Soviet countryside. The peasants were forced into new collective farms set up without adequate technical means or experienced leadership, and kulaks were forced off the land and deported in large numbers to Siberia. Their resistance was shown by their mass slaughter of their animals, which set back the production of meat and milk products for years. Soviet industry was unable to supply enough equipment, fertilizers, and insectides to raise the technical level of the new collective farms above that of the former peasant holdings, so agricultural production declined sharply, and bitter resentment long remained among the peasantry.

In industry there was certainly more enthusiasm for the change; but the emphasis of the first and subsequent plans was upon endowing the Soviet Union with heavy industry, not with increasing consumer goods production. This made high growth rates possible, since a large part of production went into new investment. Sacrifices were imposed not only on the peasantry but also upon the working class, including the huge new labour forces drained from the countryside into overcrowded industrial towns. Propaganda explained that these sacrifices would provide a better life in the future.

Social upheaval and oppression. At the price of a gigantic social upheaval the economy was transformed and entered upon a long period of expansion. The First Five-Year Plan, only unevenly fulfilled, was followed by a second in 1933 and a third still incomplete when war came. These plans maintained the priority of heavy industry, with a shift toward defense. While the capitalist world was plunged into depression, the Soviet Union grew rapidly. The growth record was impressive. Although per capita consumption may have declined and agriculture stagnated, industry—and especially heavy industry—showed a formidable advance. The development of social and community services supplemented real wages. Large numbers of people acquired new skills, as well as literacy, while a new managerial and technical elite was trained.

The Second and Third Five-Year Plans

Industrialization was achieved without foreign capital. The expected revolution in Europe had not taken place, and the isolation of the U.S.S.R. was intensified with the rise of Nazi Germany and the growth of armaments in the capitalist countries in the late 1930s. It remained economically inferior to the advanced capitalist countries despite the five-year plans.

Soviet life in the 1930s was lived under the shadow of political repression and terror unjustified by the pressures of industrialization. The arbitrary arrests, purges, trials, prison camps, and mass executions were necessary instruments for Stalin in maintaining his personal ascendancy over the new ruling stratum and the masses. But Stalinism was a burden on the economy; and it is no accident that the height of the purges is paralleled by a dip in the indexes of industrial production.

Purges and executions

Many criticisms can be made of Soviet planning. The priority given to heavy industry and armaments meant that the Soviet consumer had to suffer. Labour discipline was severe. The high output encouraged in the 1930s by the Stakhanovite Movement made for inequality of earnings. Complaints of wasteful use of productive capacity and materials and of mismanagement and corruption have frequently been made by the leaders themselves. Especially during the purges, initiative and responsibility were impaired by the fear of punishment. Bureaucratic heavy-handedness, unevenness of production and work loads made necessary by sudden spurts to achieve plan targets, production of goods not really required, and difficulties in

coordinating the requirements of different enterprises—these are among the criticisms made of Soviet planning.

The later stages of Communism. By the time Germany attacked the U.S.S.R. in June 1941 the five-year plans had given the Soviet Union a powerful industrial base. Whatever the revulsion against Stalin's crimes and the excesses and mistakes of the plans, it had been shown that an alternative to a privately-owned economy was viable. But the Soviet Union was still considerably behind the advanced capitalist countries in per capita output and labour productivity, although the gap had been narrowed. Nevertheless, without the accelerated industrialization that the plans made possible the defeat of the U.S.S.R. in 1941 or 1942 would have been certain.

At the end of the war the Soviet Union faced a tremendous task of reconstruction that again postponed the improvement in consumer goods output; but the war had spurred industrial and technological advance, and the planning apparatus was well able to meet the postwar challenge. Within a few years, however, resources were once again being diverted by the Cold War. Stalin's death in 1953 initiated a period of economic reforms. With consumer needs still generally receiving a low priority, the Soviet Union showed its strength in other directions. The sustained and rapid growth of the 1950s, the development of nuclear weapons, and finally the launching of a space satellite caused concern and even alarm in Western countries: the Soviet Union did seem to be catching up to the West.

Reforms after Stalin's death

The postwar years saw the establishment of nationalized and planned economies in eastern European countries. East Germany and Czechoslovakia apart, these countries had been predominantly agrarian, though, except for Albania, they had some industrial towns or areas. Yugoslavia broke with the Soviet Union in 1949; Albania adopted a pro-Chinese policy after the Sino-Soviet schism of the 1960s; and Romania has tended to pursue its own course. Each of the east European countries adopted the Stalinist model of "Socialism in one country" in what appeared to be an attempt to establish replicas of the Soviet Union in the "Peoples' Democracies."

The strains and tensions in Soviet society after Stalin's death were even more pronounced in a number of the east European countries. When the east European Council for Mutual Economic Assistance (CMEA) was established in 1949, much was heard about the need to develop "an international Socialist division of labour." In fact, the economic development of Czechoslovakia, Poland, Hungary, Romania, and Bulgaria had been adversely affected not only by a fundamental change in their international economic relations but also by their comparative isolation from each other. While achieving rapid rates of growth and overcoming their former underdevelopment to a considerable extent, development of these countries was uneven.

The economic history of the Soviet Union and the east European countries must be viewed against events in the capitalist world. The Cold War reduced the volume of exchanges with the capitalist countries and thus the possibility of remedying their deficiencies through trade. Moreover, scarce resources and highly trained manpower were drained away into armaments production. The easing of international tension during the late 20th century opened up the prospect of widening trade links, for which provision was made in economic plans. Growing inflation and recession abroad, however, raised the prices of sought-after imports, while making it difficult to sell the goods necessary to pay for them. Some of the price increases were passed on to consumers, and there was a tendency to run trade deficits, financed by credits from the international capital market. These difficulties were aggravated by shortfalls in the Soviet Union's harvests in some years, making necessary large grain purchases on the world market. On the other hand, as a large oil producer, the Soviet Union held a strong position vis-à-vis its trading partners in CMEA, enabling it to raise prices in line with world trends.

All these countries ran into problems in the late 20th century, and while growth was maintained, it was generally at a lower rate than in the past. Former methods of planning were revealed as overcentralized and cumbersome. Repeated attempts at reform, improving management techniques, and allowing greater play for market forces were made with indifferent success. Continued technological advance elsewhere revealed serious industrial deficiencies in fields such as computers, electronics, plastics, and petrochemicals. Great efforts were deployed to remedy such weaknesses, including the purchase of complete industrial plants from western Europe and the United States. The task of catching up with and outstripping the capitalist countries proved to be unending.

The lack of consumer goods, dissatisfaction with living conditions, and resistance to bureaucratic management produced labour problems, indicated by the low productivity often complained of in official reports and in the difficulty experienced in reaching plan targets. Rising consumer levels in western Europe had a "demonstration" effect, and the east European countries and the Soviet Union continued to be outclassed in shop-window competition. This imbalance was partially responsible for popular uprisings, but concessions to consumers continued to be limited by high armaments expenditure and the need for a high rate of investment. Consequently, automobiles and many consumer durables were available only to a privileged minority or through access to special shops. Necessities might be in adequate supply, cultural goods (such as books and records) inexpensive, social services adequate, and accommodation cheap if not abundant, but the range of goods and services the Western consumer could spend his money on was still not available.

World War II and after. The improvised character of the economic policies of the belligerents in World War I was almost entirely absent in World War II. War experience had been digested, and further lessons had been learned from the Great Depression. Perhaps most important of all, World War II had been anticipated with some deliberation by the main participants. In the end, it was the sheer weight of resources that the Allied powers were able to bring to bear that ended the war.

Economic mobilization

In the capitalist states, economic mobilization imposed unprecedented centralized control and planning. Production was regulated by a priorities system, labour was directed, and scarce commodities were rationed. Prices and wages were controlled and foreign trade was curtailed to strategic needs. The market mechanism was largely superseded, but private property rights were preserved, and, although some war factories were state owned, military contracts were generally fulfilled by private, profit-making industry. The war economy thus remained, in Germany as in Britain and the United States, a capitalist economy.

The desire to knock out the enemy's economic nerve centres and to demoralize, or even destroy, the workers upon whom the armed forces depended accounts for much of the destructiveness of World War II. Modern technology provided supremely powerful weapons that, even before the atomic bomb, could reduce whole cities to rubble in a few hours. The wide-ranging nature of the fighting, together with the effects of bombing, extended the area of destruction far more than previous wars as well as making it more complete. At the end of the war, therefore, the European countries and the Soviet Union faced an unprecedented mass of destruction. To this were added the losses from the sea war and the dislocation and disruption of prewar trade patterns.

The recovery. The recuperative powers of advanced industrial societies, however, are remarkable. The effect of bombing on machines and industrial equipment was less severe than on dwellings and other buildings. Given the human skills available and an abundant supply of people with the incentive to reshape shattered lives and earn enough to buy necessities, it was possible to restore industry and reconstruct the social infrastructure comparatively quickly. From the ruins emerged, in fact, a larger and more advanced industry than in peacetime. It was even claimed that the more devastated regions, such as western Germany, had the advantage over the less battered industries in Britain, which brought much obsolete equipment and plant intact into the postwar years. Such an explanation of British industrial problems, which were to limit its

growth in the period after 1945, was, however, too facile.

The European belligerents endured severe upheavals and hardships in the later stages of the war and the first years of peace. This was a period of general penury, of rationing and the black market, during which, in some places, money virtually lost all value, and some people returned to a barter economy. Only United States loans and aid prevented economic and social breakdown. Meanwhile, large transfers of population took place, particularly from eastern Germany and Czechoslovakia to the West, and large numbers of displaced persons had to be resettled. The many millions of rootless and penniless people provided, however, a mobile and adaptable labour force. Growth was not held back by labour scarcity, and the bargaining power of workers was held in check until full employment levels were reached.

At the end of the war a tremendous disparity existed between the United States and the dislocated economy of impoverished Europe. Only the United States could supply the foodstuffs, raw materials, semifinished products, machinery, and machine tools that were in scarce supply. The capacity of the belligerent countries to earn foreign exchange, particularly dollars, had been greatly impaired; and they needed loans or aid if reconstruction was to be carried through without balance of payments difficulties. U.S. lend-lease, under which supplies had been made available during the war, ended shortly after Japan's capitulation. Spurred by the growing tension of the Cold War, the European Recovery Program, or Marshall Plan, was proposed in 1947 (administered April 1948–December 1951) to continue reconstruction and rescue capitalism in western Europe.

The Marshall Plan

The outstanding fact was that the reconstruction boom was succeeded not by a crisis of overproduction and a return to high unemployment but by expansion that continued through the 1950s and 1960s.

Postwar monetary systems. The governments of the Allied countries in the postwar world saw that the currency chaos and beggar-my-neighbour policies of the 1930s must be avoided. First, the reconstruction of the world monetary system was needed, and the United Nations Monetary and Financial Conference held at Bretton Woods in July 1944 sought to do this through the establishment of the International Monetary Fund and the International Bank for Reconstruction and Development (World Bank). The former set out to establish monetary stability and to remove exchange restrictions that hampered world trade. It established new rules for adjusting exchange rates without a return to the old gold standard. Although each national currency was to be expressed in terms of gold, it was the dollar price of gold and the convertibility of the dollar that together provided the essence of the postwar monetary system.

The recovery of the world market was much more rapid than after World War I. In an atmosphere of expansion, international trade was liberalized through the General Agreement on Tariffs and Trade (GATT), and other agencies made possible a greater degree of international economic cooperation. Some, such as the Organisation for European Economic Cooperation (OEEC), were mainly consultative bodies. More important was the movement toward economic integration in Europe that led to the Treaty of Rome in 1957 and the setting up of the European Economic Community (EEC, or the Common Market) in 1957.

Economic expansion. From the early 1950s a long-term expansion began in the industrialized countries. Particularly rapid growth rates were achieved by advanced countries that were incompletely industrialized. Western Germany resumed its place as the hub of industrial northwest Europe. The Scandinavian countries continued with their solid prosperity. The fringe countries of southern Europe, still relatively backward, began to speed into the 20th century. Such older industrial areas as Britain and Belgium experienced growth, too, but at a rate that compared unfavourably with that of the rest of western Europe.

The roots of the sustained growth of the advanced industrial countries in the 1950s and 1960s seem to have been laid in the war and even as far back as the Depression. A long period in which investment was at a low level and, in some fields, actually negative had cleared the way. In some countries the war had added to industrial capacity; machines and machine tools were available to produce new consumer goods. Despite the destruction in Germany, there was more plant and equipment available for production than in the 1930s. British industry, which had not entirely thrown off the Depression in 1939, entered the postwar world in a sellers' market, and unemployment was minimal until the late 1960s.

Background of postwar expansion

Interrelated developments in varied fields tended to make industry much more science based than ever before. This enabled the logic in the industrial system to be worked out more fully, culminating in the fully automated computer-controlled process. But highly sophisticated machines had only a partial application and represented an enormous outlay of capital. Perhaps more important in raising total output was the steady increase of productivity levels over almost the whole of industry. Into this there entered not only straight technology but also changes in business organization, administration, and management, and a better educated and more adaptable labour force.

Postwar capitalism was characterized by almost uninterrupted output and productivity growth and high levels of investment. As the expansion continued its momentum, workers had to be mobilized from further afield. Germany and France each attracted some 3,000,000 immigrants to meet a chronic labour shortage. In particular, steady and prolonged employment became a part of the expected order of things. Full employment greatly enhanced trade-union power as well as making employers compete for labour. Some economists claimed that "wage push" was a main factor in the inflationary pressures of the late 20th century; but the whole situation was inherently inflationary, and governments that attempted to deal with it merely checked expansion and created a recession. It was generally accepted, therefore, that some inflation was unavoidable if growth and full employment were to be maintained. Government expenditures, especially on armaments, when financed by credit expansion, fed the inflationary trend and were probably its basic source. The ability to increase the money supply was made possible by the currency arrangements established by Bretton Woods and particularly by the outflow of dollars from the United States.

Internationalism. The European Economic Community, or Common Market, was in the first place a customs union whose members reduced (and were finally to abolish) duties on their own trade while a common external tariff was to be maintained against imports. The original signatories of the Treaty of Rome—France, West Germany, Italy, Belgium, The Netherlands, and Luxembourg—also intended to develop policies to achieve full economic integration. It was hoped that the establishment of a unified market area would yield benefits comparable to those gained by the United States. Great Britain, Denmark, and Ireland joined the Common Market in 1973, and Greece in 1981.

The Treaty of Rome

The giant international firm was another focal point of public interest. These firms were particularly strong in the industries that were in the vanguard of technical progress, such as computers, motor vehicles, and electronics.

As incomes in Europe rose, more was spent on luxury goods; and the extent to which a country was equipped with different types of such goods became a reliable index of its per capita income. Even advanced European countries were poorly equipped with washing machines, refrigerators, vacuum cleaners, and the like after World War II, and thus a big field was open for new investment and large-scale production, with U.S. capital often taking a leading part. As a result, European consumption moved closer to U.S. patterns. Europeans also spent more on services of various kinds, so that there was a continuous growth of the "tertiary" sector in Europe and industrialization itself resulted in a fall in the proportion of the occupied population working in manufacturing industry.

The uncertain future. By the early 1970s it was clear that the long-sustained postwar expansion had come to an end. Most countries were subject to higher inflation

rates coupled with slower, or arrested, growth—a situation described by some economists as "stagflation"—and unemployment reappeared as a serious problem. Heavy industries such as steel and shipbuilding were hardest hit as investment rates declined, as were some traditional consumer-goods industries subject to sharper foreign competition. What was spoken of as a "crisis" worsened with the decision of the Organization of Petroleum Exporting Countries (OPEC) to increase prices fourfold following the Arab–Israeli War of 1973. Importing countries were faced with the problem of how to meet the higher payments, while the oil producers acquired large liquid reserves, mainly in dollars, the movements of which imposed further strain on the already dislocated international monetary system. Continued inflation tended to reduce the real cost of oil imports, and OPEC replied with further price increases. The oil-importing countries that had strong exports responded by increasing their foreign earnings to cover the higher import bill. Oil-importing countries with weaker economies suffered balance of payments deficits that weakened their currencies and drove them into debt.

During the late 20th century the governments of the economically advanced countries agreed that the struggle against inflation would have to take priority over the maintenance of full employment. Clear-cut policies were slow to take shape, but there was no doubt that the influence of Keynesian economics and the use of its remedies against recession had waned. In general, government withdrawal took the form of abstention from new measures of intervention to meet recession or deal with unemployment and some dismantling of existing controls. These policies met with opposition, especially from organized labour, whose cartel-like methods in wage bargaining were also under attack. Economic issues bound up with inflation and unemployment were very much at the centre of political debate and were likely to remain so in the late 20th century. (T.Ke./Ed.)

MODERN CULTURE

In the last quarter of the 19th century European thought and art became infused with doubt and decadence. Writers as different as Baudelaire and Matthew Arnold, Henry Adams and Flaubert, Ruskin and Nietzsche were exasperated by the banality and smugness of surrounding humanity. It seemed as if with the onset of Positivism and science, *Realpolitik* and Darwinism, realistic art and popular culture, all noble thought and true emotion had been suffocated. The only things that stood out from banality and smugness were their own appalling extremes—vulgarity and arrogance—against which all the weapons of the mind seemed powerless.

The diminished influence of the intellectuals

Such intellectuals and artists were hopelessly outnumbered not only in the literal sense but also in the means of influencing culture. A newspaper that reached half a million readers with its cliches, its serial story, and its garish illustrations "educated" the people in a fashion that actively prevented any understanding of any part of true culture. The barrier was far more insurmountable than mere ignorance or illiteracy; and it was cutting off not just the populace but also—to use Arnold's terms—the barbarian upper class and the Philistine middle class. Similarly, Nietzsche anatomized the culture Philistine; that is, the person whose mind was being filled by the mediocre thoughts of writers and artists high in repute and generally believed to be leaders of civilization. In other words, the prudent, self-limiting impulse of Realism after 1848 had generated the middlebrow, while the evolution of industrial democracy had generated the mass man. By the late 1880s the gap between this compact army with its honoured officers and common soldiers and the hostile, half-visible avant-garde was a permanent feature of cultural evolution.

The individual expressions of disgust with the former age of Realism naturally took many forms, but they can be grouped into half a dozen kinds, not all on the same plane and only some of which have come to bear distinctive names. Some artists retreated from the ugly world into a species of Neoclassicism. Such were the French poets known as Parnassiens. Strict form, antique subjects, and the pose of impassivity constitute their hallmark. In painting, the work of Puvis de Chavannes stands in parallel.

In music, the explicit revolt against Wagner and Liszt, of which Brahms was made the torchbearer, offers similarities, particularly in the desire to learn and employ the "purer" forms of an earlier time. Likewise, the shift in tone and temper of the later poems of Tennyson, Arnold, or Gautier; the resurgence of Thomist orthodoxy in Catholic thought; the haughty detachment in the plays of Becque and those of Ibsen's middle period, all suggest a search for stability, for a fixed point from which to judge and condemn contemporary "progress."

Symbolism and Impressionism. It was also possible, after withdrawing by conscious effort from any struggle with vulgar actuality, to make art itself an independent and inhabitable world and to escape alienation in beauty. This cult of beauty goes by the names of Symbolism and Impressionism. The Impressionists were able to take as subjects some of the sights that most depressed their fellow man and by recomposing them in brilliant, shimmering colour to create a refreshing world of new sensation. Subject once again mercifully disappeared. As Monet said: "The principal subject in a painting is light."

The Symbolists in literature had a more difficult task than the painters, because their medium, words, must be shared with all those who speak the language for ordinary purposes. To disinfect grammar and vocabulary for poetry and "art prose" required severe measures. All set phrases had to be broken up, unusual words revived or common ones used in archaic or etymological senses; syntax had to be bent to permit fresh juxtapositions from which new meanings might emerge; above all, the familiar rhetoric and rhythms had to be avoided, until the literary work, poetry or prose, created the desired "new world." It is a world difficult to access but worth exploring, all its tangible parts being the symbols of a radiant reality beyond—in short, the antithesis of a newspaper editorial.

In music there was no need of any indirect device to establish the mood of Impressionism. It was already to be found here and there in the great Romantics, and when the new generation began to compose on themes drawn from literature, the hints and opportunities needed only a delicate genius to develop them into a style. Debussy was that genius, soon followed by Ravel, Delius, Hugo Wolf, and others. Alike, yet independently of one another, they replaced eloquence, melodic clarity, and harmonic consecutiveness by capricious melodic contour and pointillist chord progressions to produce the shimmer and mystery of musical Impressionism.

Aestheticism. To those who dedicated their lives to Symbolist literature and criticism the name of aesthetes is often given, for it was at this time, from 1870 to the end of the century, that questions of aesthetics became the intense concern of artists, critics, and a portion of the public. The old Romanticist phrase "art for art's sake" was revived; many persons not themselves artists confessed to a "cult of art"; much time was devoted—necessarily—to deciphering, commenting, and discriminating, for much of the new art was obscure and arcane and took time to master and enjoy. Other observers called these manifestations decadent—some with contempt, others with pride. For a time, a group of French artists took the name to themselves as a badge of honour.

But aestheticism was by no means as languid and fatalistic as it tried to appear. At its centre stands the active figure of Oscar Wilde, whom it is easy, because of his notoriety on many counts, to dismiss as colourful but ephemeral. Yet he is far more than a symptomatic figure; he is a representative one, and a good part of his work, namely his entire critical output, some of his fiction, one or two poems, the autobiographical *De Profundis,* and the greatest farce in the language, *The Importance of Being Earnest,* deserves to last. What Wilde accomplished through these works was the liberation of English literature from ancestral (and not merely Victorian) preconceptions. He reconnected England with the Continent artistically by phrasing with finality their different assumptions. He showed that art could be morally responsible only by discarding moralism. In a word, he played again in 1890

the role Gautier had played in France in 1835 with his art for art's sake diatribe in *Mademoiselle de Maupin.* Whoever, starting with Wilde or Gautier, wishes to follow the historical sequence and recapture the atmosphere in which this activity went on will find no better source than the journal of the Goncourt brothers, the inventors of "art prose" as well as the recorders of contemporary lives, characters, and gossip.

But the reader of their voluminous pages will also find there references to the movement called Naturalism, which does not merely parallel but also intermingles with Symbolism and Impressionism. The Goncourts themselves wrote a number of Naturalistic novels; their friend Zola was the theorist and greatest master of the genre; another novelist, J.-K. Huysmans, passed from Naturalism to Symbolism, as did several other writers. In the poets Rimbaud and Verlaine, as later in the Irish Yeats, the elements of the two tendencies alternated or mixed.

Naturalism. The name Naturalism suggests the philosophy of science, and the connection is genuine. Zola thought that in his great series of novels, *Les Rougon-Macquart,* he was studying the "natural and social history" of a family during the time of Napoleon III. The claim was bolstered by the method Zola used of gathering data like a scientist—every material fact could be proved by reference to actuality or statistics. Naturalism would thus appear to be an intensification of Realism, as indeed it was—more "research." But it differed markedly in spirit. Realism professed to be depiction of the commonplace in a mood of stoicism or indifference—a photographic plate from a camera held almost at random in front of unselected mediocrity; it was, as Flaubert was the first to say, a refusal to share previous Romanticist hopes and interests. Naturalism, on the contrary, readmitted purpose and selectivity. Each novel was a "study" designed to show up and denounce the dismal truths of social existence, for which purpose the worst are the best. Zola's novels throb with a passionate love of life, a life which he showed as tortured and twisted by character and condition. In the end he defined his scientific or "experimental" novel as "nature seen through a temperament."

In the plastic arts, a plausible counterpart of Naturalism is the work of those known as Postimpressionists, notably Cézanne and van Gogh in painting, Rodin and Maillol in sculpture. Their various styles and aims had a common result in restoring solidity and "weight" to the visual object after the fluidity and lightness of Impressionism.

Musical naturalism was, by contrast, an attempt at dramatic literalness. Richard Strauss boasted that he could render a soup spoon in music if he wanted to; but he could not and he did not. The noises of his *Sinfonia Domestica* are standard orchestral sounds fitted with a preliminary explanation, like the libretto or synopsis of a Wagnerian or other opera. When the sheep bleat in Strauss's *Don Quixote,* the clarinets play notes that are decorative on their own account and do not in the least suggest wool. It is rather the thickness of Strauss's orchestration and chromatic harmony that connect him with naturalist doctrine—the headlong embrace with matter. And so it is also in the operas of Bruneau or Charpentier or in the *verismo* of Puccini and the late Italian school generally. Music remains atmospheric; never, except in Wagner's system, denotative.

This definition of Naturalism, coupled with the aesthetic, or "art for art's sake," impetus in Symbolism and with the Impressionists' transmutation of concreteness into light, justifies the name of Neo-romanticism that has been given to the cultural temper with which the 19th century ended. After the glum self-repression of the middle period, it was an outburst of vehement self-assertion, whether directed inward or outward. "Art for art's sake" and Naturalism are indeed but twin branches of one doctrine: art for life's sake.

The new century. In 1895 George Bernard Shaw said: "France is certainly decadent if she thinks she is." The remark is characteristic of Shaw, but also symptomatic of a new wave of energy. From under the despair and decadence, the scattered retreats and the violent nihilism, the same human strength that produced Symbolist and Naturalist art was trying to reshape the civilization that all found so unsatisfactory.

In England, the Fabians, of whom Shaw was one, were preaching the "inevitableness of gradualism" toward the socialist state. It was they, seconded by the growing strength of the trade unions after a spectacular dock strike of 1889, who paved the way to Labour governments and the British welfare state. Throughout Europe, Socialism was no longer the creed of a lunatic fringe but was the ideal of many among the masses and the intellectuals. The original fight for liberty and democracy in political action had turned into a fight for economic democracy—freedom from want. Laissez faire liberalism had turned inside out, and the liberal imagination at work in the many brands of socialism now demanded state interference to remove the appalling conditions causing all the despair.

The growth of Socialism

Arts and Crafts movement. Among the Socialists belonging to no party, Ruskin and William Morris worked also to effect immediate changes in the quality of their surroundings: they started the so-called Arts and Crafts movement, whose aim was to make objects once again beautiful. Because machine industry produced only the cheap and nasty, they tried by hand to produce the cheap and handsome—good furniture, hangings, and household articles; fast dyes of good colour; well-printed books on good paper; and jewelry and ornaments of all kinds that showed visual talent as well as manual skill. In a word, the movement reinstated the ideal of design and succeeded in forcing it on machine industry itself. Within two decades manufacturers began to hire artists as designers, and by 1910 the 20th-century omnipresence of design, from clothes to print and from gadgetry to packaging, was a *fait accompli.* The visual revolution can be seen easily by looking back with modern eyes to a page of advertising at the turn of the century.

New trends in technology and science. In parallel with the new craftsmanship, the new technology of the 1900s began to give hope of wider improvements. The use and transmission of electric power suggested the possibility of the clean factory, all glass and white tile. Better machines, new materials and alloys, a greatly expanded chemical industry—all supplied more exact, more functional, less hazardous objects of use and consumption, while the application of science to medicine nourished the hope of greatly reducing the physical sores of mankind. Those closest to all these developments were certainly not among the despairers and fugitives from the world. Like all men who struggle successfully with difficulties, they were inspirited by what they knew to be demonstrable progress along their chosen lines.

The same outlook animated workers in the natural and social sciences. It was for both a time of transformation, and genuine novelty exerted its usual invigorating effect. From the '80s onward it had been clear that simple mechanistic explanations based on "dead" matter were inadequate. The Michelson-Morley experiment of 1887 had given the *coup de grace* to the mere push-pull principle by showing that, though light consisted of waves, the waves were not in or of anything, such as the ether, which did not exist. Even earlier, James Clerk Maxwell's attempt to work out the facts of electromagnetism on Newtonian principles had failed. And on the philosophic front, the notion of natural "laws" was being radically modified by thinkers such as Poincaré, Boutroux, Ernst Mach, Bergson, and William James. All this prepared the ground for the twin revolutions of relativity and quantum theory on which the 20th-century scientific regime is established.

The decline of the machine analogy had its counterpart in the biological sciences. With narrow Darwinian dogmas in abeyance, the genetics of Gregor Mendel were rediscovered, and a new science was born. The fixity of species was again regarded as important (Bateson), while the phenomenon of large mutations (de Vries) caught the public imagination, as the opposite slow, small changes had done 60 years earlier. The mysterious "fitness of the environment" was held deserving of study, and new formulas for evolution—emergent, creative—received serious attention. Vitalism once more reasserted its claims, as it seems bound to do in an eternal seesaw with mechanism.

Rediscovery of genetics

The social sciences. Finally, in the social sciences, fresh starts were made on new premises. Anthropology dropped its concern with physique and race and turned to "culture" as the proper unit of scientific study. Similarly in sociology, Durkheim, seconded by Tönnies, Tarde, and Le Bon, concentrated on "the social fact" as an independent and measurable reality equivalent to a physical datum. Psychology, also long under the exclusive sway of physics and physiology, now established at the hands of William James that the irreducible element of its subject matter was the "stream of consciousness"—not a compound of atomized "ideas" or "impressions" or "mind-stuff" but a live force in which image and feeling, subconscious drive and purposive interest, were not separable except abstractly. A last domain of research was mythology, to the existence of which James George Frazer's *Golden Bough* gave massive witness, thereby exerting proportional influence on literature and criticism.

Reexamination of the universe. The net effect of these innovations in the sciences of man and of nature was liberating. Whatever each specialty or subspecialty meant to its practitioners, the persons who carry in their minds the general culture of an age took the new message to mean that the universe, formerly closed forever, had been reopened, and that on fresh inspection it had proved more alive than dead. These conclusions offered a field for strong individual action. Subjective choice became again respectable as a genuine fact. Truth, God, beauty, and immortality obviously were in need of reexamination.

In philosophy, politics, and criticism this reexamination may be called the pragmatic revolution; in social and moral life, the liquidation of Victorianism. But the Pragmatic Revolution must not be thought of as being only the work of those who, like James, called themselves Pragmatists. Nietzsche, Samuel Butler, Shaw, Bergson, and others constitute the headwaters of the stream of thought that issues in present-day Existentialism. The common features are the turning away from absolutes and unities to pluralisms and the method of testing by consequences. Subjective and objective tests looking to future thought and action—not authority or antecedents—are to decide the true, the good, and the beautiful.

Turning away from absolutes

Such an outlook, of which the refinements are, like the defects, beyond the scope of this article, is the logical and appropriate one for an age of reconstruction. It boils down to trying all things new and holding fast to that which is good. But it presupposes the creation of new things to try, and here it is allied to the liquidation of Victorianism. In morals the work of destruction generally begins by affirming the opposite of the accepted rule. An excellent source book for this attitude is Samuel Butler's *Way of All Flesh,* written in 1885 but not published till 1903. The Victorian Tennyson had said: "'Tis better to have loved and lost than never to have loved at all." Butler said: "'Tis better to have loved and lost than never to have lost at all." This inversion of values—don't weep over loss; there are plenty of loves to be had and the more the merrier—is but an indication of method. At first the denial was uttered as humour and paradox: Butler's *Note-books,* Shaw's *Arms and the Man* (the soldier wants chocolate, not ammunition), Wilde's *Importance of Being Earnest,* Jarry's *Ubu roi,* Strindberg's tragicomedies—to cite but a few subverters of the Victorian—all used derision and topsy-turviness to make their point.

But underneath the joke was the new purpose, which soon found open expression in positive utterance and action. In the plays of Hauptmann and Brieux, the novels and anticipations of H.G. Wells, the essays of Tolstoy, Péguy, Georges Sorel, Ellen Key, Havelock Ellis, Unamuno, Ortega y Gasset, or Shaw, the new modes of feeling and the new scale of virtues and vices are set forth with as much earnestness and vigour as the old Victorian kind.

Nor did action wait until all the books were out. From the onset of the overturn, say 1890 onward, the rebellion was a biographical fact. People braved public opinion and got divorced, lived together unmarried, practiced and preached contraception, studied the psychology of sex, and defended homosexuality. Or again, the sons of the rich turned Socialist, became labour leaders, and fomented syndicalist (*i.e.,* direct action) strikes, while the daughters demanded the vote as suffragettes, assaulted policemen, and went to jail for chaining themselves to door handles. Meanwhile, students rioted about international incidents or university affairs; schools were subjected to the devastation of the softer pedagogies; "rational clothing" exhibited itself in spite of derision, like the bicycle and the newfangled automobile; and new cults multiplied like mushrooms—outdoor sports, nudism, Theosophy, esoteric Buddhism, Rosicrucianism, New Thought, the Society for Psychical Research, Christian Science, the Salvation Army, and the "Maximinism" of Stefan George.

Of these, hardly any need explanation here. But a word must be added about Theosophy if only because of its historical importance in developing Yeats's genius and for expressing once again the attraction that the "wisdom of the East" has for Westerners. Not that the doctrine elaborated by Madame Blavatsky rested on any genuine knowledge of Hindu religion and philosophy. That is not its point. The point is rather that Theosophy supplied the need for quietude, mystery, transcendence, and immortality in the wearied souls of Europeans. In Theosophy the doctrine of reincarnation offers satisfaction of immortal longings and inspires to wisdom, the demands of which are periodically revealed by mahatmas, or holy men.

Theosophy

As for the poet Stefan George's worship of his young friend Maximin, who died at 15, it answers a similar impulse to permanent truth but with the additional urge to abolish (rather than escape) "contemporary materialism." George was but one among many European writers who wanted to found a new society in place of the actual one. What has fitly been called the politics of cultural despair fastened on a great many saviours as the new hope—monarchy, "integral nationalism," a new aristocracy (usually tinged with intellect), technocracy (rule by science and the engineers), the proletariat (in syndicalist "cells" or communist collectives), trade and professional guilds federated in a corporate state, or again the mystic unity of "blood" and "race." In all these creeds, at least at their beginnings, the thirst for the ideal is evident: together they formed a new utopianism, of which the later fruits are familiar but quite other than those predicted: Soviet and Chinese Communism, Italian Fascism, German National Socialism. Beyond them, in the present, elements of this second wave of cultist thought and practical utopianism that ushered in our century are easily discerned.

In one country, as the 19th century passed into the 20th, all the energies seeking an ideal found an unexpected outlet. The occasion for battle was the conviction of a French officer for espionage; *i.e.,* the Dreyfus affair. Its cultural suggestiveness is apparent: on one side, the ideal of justice and the regard for the individual as an end in himself; on the other, the social or collective ideal typified by the army and the nation; throughout, the ideal of truth—the facts—pursued, lost, and found again in an embittered struggle that threw up a host of endemic prejudices—about race, about class, for and against intellect—to say nothing of individual egotisms and obsessions that had been charged with the force of pent-up aggression.

The prewar period. The same universal aggressiveness was to have its field day in the coming war of nations, but in the intervening decade (1905–14) occurred the remarkable outburst of a creativeness, which, for the first time since 1789, had its source elsewhere than in Romanticism. The "Cubist decade" (as it has been conveniently called) gave the models and the methods of a new art, just as the natural and social sciences had begun to do for science a little earlier. Cubism in painting defined itself as a new classicism, but it was obviously not Neoclassical. In painting and sculpture, in music and poetry, and in architecture especially, the new qualities were: simplicity, abstraction, and the importance of mass.

The Cubist decade

Since the artists at work were numerous and diverse, in both origins and purpose, it is idle to try to derive their productions from any single set of conditions. Clearly, the Cubist painters found encouragement in the canvases and ideas of their immediate predecessors, the Postimpressionists. But the new architects found it rather in new materials—reinforced concrete and the steel constructions of the

engineers. Some of the poets (Futurists, Unanimists, Simultanists) were influenced by sociological theories about the group; others (and novelists too) read the new psychologists and philosophers. The sculptors responded to African and Javanese artifacts. And the musicians perceived at once the exhaustion of a system of harmony, the impact of the crowd and the primitive, and the opportunities offered by using and making sounds hitherto ignored by composers. In a culture as cosmopolitan and quick-witted as that of Europe before 1914, it is unlikely that any important concept or speculation should not have found its mark somewhere in art, even if the individual artist did not consciously record it.

Consequently, it is not arbitrary to say that the Cubist decade was stimulated and prepared for its characteristic work by everything that was pouring in upon it—motion pictures and the airplane, X rays and organ transplants, the philosophy of time and space in relativity and Bergsonism, the wireless and the Eiffel Tower, the Jamesian "stream of consciousness" and Le Bon's analyses of the crowd, the importations of colonial art and the European discovery of Walt Whitman, the myths in Frazer and the double personalities in the psychopathic clinics, the fever of revolution and of sexuality reacknowledged, the transcendental visions of various orientalisms and the vitalist current that was reanimating the cosmos and authorizing an infinite number of geometries. Multiplicity outside, bottomless depths within, equally validated by science and intuition, furnished unlimited scope to perceptive genius; it was only a question of pondering, choosing, and creating fit forms.

The impact of war. To the many-sided productiveness the war of 1914 put an instantaneous stop. It was a war of a sort Europe had not known since 1815—the nation in arms. And at that earlier time, the absence of large industry had precluded the involvement, physical and mental, of every adult citizen simultaneously throughout Europe. In 1914 Beethoven, Hegel, and Goethe would have been in the trenches.

The death of Renaissance high civilization

The cessation of cultural activities; their replacement everywhere by a propaganda of hate; the rapid decimation of talent and genius in the murderous warfare of bombardment and infantry assault; the gradual demoralization through four years of less and less intelligible war aims; and after the Armistice, the long sequel of horrors—starvation, dispersion, disease, and massacre—together shattered the high civilization born of the Renaissance and based on the idea of the national state. Too many able men and women had been killed at the start or at the peak of their creative powers for the continuity of culture. Too many intimate faiths and civil traditions had been ground down for any recovery of self-confidence and public hope to be possible. Skepticism about this conclusion is easily met by reference to the present-day view of "the human condition" as it is reflected in novels and plays.

The period "between wars"—1920 to 1939—gave sufficient indications that forward direction had been lost. It combined a frivolous neo-decadence (if the term may be allowed) with a restless search for starting points. The artists who survived were unable to pursue their individual development from the point of its interruption, whether physical or spiritual. Many felt a kind of shame to be busy with art after the human disaster they had witnessed. This feeling accounts for the tiresome mocking of oneself and one's work that marks the production of the 1920s and '30s. Nothing being important, everything is silly, and it can be offered at all only if one laughs at it first. Hence the appearance of Dada, the meaning of which is that the thing is a harmless and even childish obsession.

More serious minds—Stravinsky's, for example—looked farther back than prewar times for possible models: the 18th century or the late Middle Ages. Playwrights took up the ready-made of antiquity, like 17th-century "ancients," but turned the plots inside out or upside down, those acrobatics mirroring their times and their souls simultaneously. T.S. Eliot, after expressing his disgust at all he saw and felt, declared himself a royalist, an Anglo-Catholic, and a Neoclassicist. In all the arts, eclecticism and, even more, allusiveness were the preferred methods. The partakers' enjoyment came from recognizing the borrowed plums. To this device the sole alternative was to deny the validity of art itself, as in Surrealism, which relied for its deliverances on automatic writing uncensored by the conscious mind. The ultimate creator was believed to be the dark side of our common nature.

Culture since 1920. It is always true, of course, that as the observer comes nearer to his own times his blindness increases. When he has said his say, it turns out that something or someone remarkable has been overlooked—Blake or Kierkegaard, the aesthetics of Poe and Whitman, or the scientific truths of Mendel or Semmelweis—all delayed by the myopia of the ignorant and the well informed. Therefore, any generalities made about the period since 1920 are and ought to remain suspect for a good while. Still, they are not without value for the present for two reasons: they are suggestive and they allow for a sifting through of a variety of tenable views. In addition, such conclusions inform later comers of how things seemed before time had cleared away the rubble.

The most important truth about culture since 1920 is the so-called cultural lag. The meaning of the term is simple: culture in any sense at all is being made somewhere and it takes five, 10, or 20 years to become known and diffused. The contributions of Freud and Einstein, for example, did not enter the public domain until 1920, though clearly perceivable by 1905. Subtracting five years for war and peace leaves a net lag of a whole decade. Again, there is the North Atlantic cultural lag. Until the first war sent American youth to Europe and conscious expatriation followed, the lag between the continents was 20 years or more. But since 1920, in both Europe and America, a different kind of cultural lag has been noticeable. It is the joint result of swift social evolution and failure to move creatively beyond the Cubist decade.

The "cultural lag"

The social evolution has consisted in bringing new layers of the population in contact with cultural affairs. The mass media have done the work of popularizing—radio, television, and magazines have reached the millions and told them in words and pictures "all about art"; while display advertising has quietly adapted the tone and techniques of modern literature and painting to charm and seduce the customer. The schools, including the universities, have likewise "gone modern" and bred in their students a taste for novelty in place of the classics. It could be argued that for 80 years or more, the preponderance of effort has been aimed at making an end of the aristocratic, elitist high culture of Renaissance man and his successors. That that culture has democratized itself and brought to the fore all the ideas of liberty, equality, social responsibility, humaneness, world unity, and cosmic brotherhood has not helped to save it.

In such a state of affairs, only one supposition may be possible: European man has thrived for half a millennium on the principle of individualism and has exhausted all its possibilities. A new type of man, actuated by a different principle, is required to cope with the conditions created by his predecessor. If he arises, a new culture—art, science, philosophy, social and moral relations, and political state—will follow. (J.Ba.)

BIBLIOGRAPHY

Ancient history: J. BOHM and S.J. DE LAET (eds.), *L'Europe à la fin de l'âge de pierre* (1961), numerous articles by specialists covering the later Neolithic settlement of Europe, mainly on central Europe and the Balkans; V.G. CHILDE, *The Dawn of European Civilization* (1957), the last (6th) edition of Gordon Childe's classic work, still the best survey of European prehistory from the Mesolithic to the Early Bronze Age; G. CLARK, *World Prehistory: A New Outline,* ch. 6–7 (1969), a description of the settled communities of prehistoric Europe before and after *c.* 1500 BC, and with S. PIGGOTT, *Prehistoric Societies,* ch. 10–13 (1970), on the peasant societies of prehistoric Europe; J.G.D. CLARK, *Prehistoric Europe: The Economic Basis* (1952), on the subsistence economies and technology of prehistoric Europeans from the Neolithic to Iron Age; M. GIMBUTAS, *Bronze Age Cultures in Central and Eastern Europe* (1965), especially valuable as a summary of sources originally published in Russian and other eastern European languages; S. PIGGOTT, *Ancient Europe* (1965), aspects of the cultural history of prehistoric Europe from Neolithic times to the Celtic period; N.K. SANDARS, *Prehistoric Art in Europe* (1968), especially valuable for Neolithic Balkans and for La Tène art.

Middle Ages: The fullest account of medieval Europe in English is that of the *Cambridge Medieval History,* 8 vol. (1924–36; 2nd ed., 1966–), of which only the Byzantine volumes have been re-edited, as has the first volume of the companion *Cambridge Economic History of Europe* (1966). GEORGES DUBY, *L'Économie rurale et la vie des campagnes dans l'occident médiéval,* 2 vol. (1962; Eng. trans., *Rural Economy and Country Life in the Medieval West,* 1968), summarizes much recent work and shows well the directions of current research; FRANCOIS L. GANSHOF, *Qu'est-ce que la féodalité?,* 2nd ed. (1947; Eng. trans., *Feudalism,* 1952), is the classic analysis of the structures of feudal tenure. FRITZ KERN, *Gottesgnadentum und Widerstandsrecht im früheren mittelalter* (1914; Eng. trans., *Kingship and Law in the Middle Ages,* 1939, reprinted 1968), is the most convenient introduction to its subject. *The Handbook of Church History,* ed. by HUBERT JEDIN and JOHN DOLAN (1965–), gives a full and recent account of the period from the 8th to the 12th centuries; for the later period, A.C. FLICK, *The Decline of the Medieval Church* (1930), is still useful; R.W. SOUTHERN, *Western Society and the Church in the Middle Ages* (1970), gives a more general introduction. On the towns HENRI PIRENNE, *Medieval Cities* (1925, reprinted 1956), has lost little of its value; DANIEL P. WALEY, *The Italian City-Republics* (1969), covers the period between the 11th and the 14th centuries in greater detail. GEOFFREY BARRACLOUGH (trans.), *Medieval Germany, 911–1250: Essays by German Historians* (1961); JOHANNES HALLER, *Die Epochen der deutschen Geschichte* (1928; Eng. trans., *The Epochs of German History,* 1930); ROBERT FAWTIER, *Les Capétiens et la France* (1942; Eng. trans., *The Capetian Kings of France,* 1960); and EDOUARD PERROY, *La Guerre de cent ans* (1945; Eng. trans., *The Hundred Years War,* 1951), give useful examples of national history. On England the first 5 volumes of the *Oxford History of England* (1937–59), are authoritative. JOHAN HUIZINGA, *Herfsttij der Middeleeuwen* (1950; Eng. trans., *The Waning of the Middle Ages,* 1954), is a controversial classic. The largest and most representative collection of medieval texts in translation is that in the Columbia "Records of Civilization" series, supplemented by that of the *Oxford* (formerly *Nelson*) *Medieval Texts.* The standard English bibliography of medieval history is L.J. PAETOW, *A Guide to the Study of Medieval History,* rev. ed. (1931, reprinted 1964). See also P.H. SAWYER, *Kings and Vikings: Scandinavia and Europe, A.D. 700–1100* (1983).

History, 1500–1648: The following general works include their own bibliographies: *The New Cambridge Modern History:* vol. 1, *The Renaissance, 1493–1520;* vol. 2, *The Reformation, 1520–1559;* vol. 3, *The Counter-Reformation and Price Revolution, 1559–1610;* and vol. 4, *The Decline of Spain and the Thirty Years War, 1609–1648/59* (1957–70), the best recent textbook of general history written in English; *Handbuch der mittelalterlichen und neueren Geschichte,* vol. 2, *Geschichte des europäischen Staatensystems von 1492–1559* (1919; Italian trans., 1932); WALTER PLATZHOFF, *Geschichte des europäischen Staaten-Systems, 1559–1660* (1928, reprinted 1968), a concise, useful account of international relations; GASTON ZELLER, *Les Temps modernes,* vol. 1, *De Christophe Colomb à Cromwell* (1953), a critical and stimulating study; and HENRI HAUSER, *La Prépondérance espagnole, 1559–1660* (1933, reprinted 1972), most illuminating. ERIC R. WOLF, *Europe and the People Without History* (1982), is a cultural history since 1400.

Renaissance: HANS BARON, *The Crisis of the Early Italian Renaissance,* 2nd ed. (1966), a brilliant interpretation of the rise of civic Humanism; R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954), a useful survey of classical scholarship and education; ERNST CASSIRER, *Individuum und Kosmos in der Philosophie der Renaissance,* 2nd ed. (1963; Eng. trans., *The Individual and the Cosmos in Renaissance Philosophy,* 1963), not confined to Italy—the central figure is the German Nicholas of Cusa; EUGENIO GARIN, *L'umanesimo italiano,* 2nd ed. (1958; Eng. trans., *Italian Humanism,* 1965), a major interpretation of Renaissance Humanism by a leading historian; FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965), chiefly concerned with the two Florentines as historians but contains a great deal more on the historical context of their lives and thought; PAUL OSKAR KRISTELLER, *Renaissance Thought,* 2 vol. (1961–65), the best introduction to the work of a leading historian of Renaissance Humanism and philosophy; GINO LUZZATTO, *Breve storia economica d'Italia: dalla caduta dell'Impero romano al principio del cinquecento* (1958; Eng. trans., *An Economic History of Italy: From the Fall of the Roman Empire to the Beginning of the Sixteenth Century,* 1961), an excellent survey; GARRETT MATTINGLY, *Renaissance Diplomacy* (1955), the best introduction to the subject by a brilliant historian and writer; ERWIN PANOFSKY, *Renaissance and Renascences in Western Art,* 2nd ed., 2 vol. (1965), an eminent art historian on the distinctive nature of Italian Renaissance art; DANIEL P. WALEY, *The Italian City-Republics* (1969), the best introduction to the medieval development of Italian city-states; J.H. WHITFIELD, *Petrarch and the Renascence* (1943), an excellent study by a leading historian of Renaissance literature; PHILIP ZIEGLER, *The Black Death* (1969), a good popular account. ROLAND H. BAINTON, *Erasmus of Christendom* (1969), one of the most recent and readable of the many books on Erasmus; OTTO BENESCH, *The Art of the Renaissance in Northern Europe,* rev. ed. (1965), a major study by a leading scholar; KARL BRANDI, *Kaiser Karl V,* 2 vol. (1937–41; Eng. trans., *The Emperor Charles V,* 1939), a superb biography of a pivotal 16th-century figure; J.M. CLARK, *The Great German Mystics: Eckhart, Tauler and Suso* (1949), a good introduction to the three founders of the German school; J.H. ELLIOTT, *Imperial Spain, 1469–1716* (1963), the best short account in English; H.A. MISKIMIN, *The Economy of Early Renaissance Europe, 1300–1460* (1969), a brief, up-to-date survey; E. ALLISON PEERS, *Studies of the Spanish Mystics,* 3 vol. (1927–60), the classic work in the field; GUSTAVE REESE, *Music in the Renaissance* (1954), an excellent, comprehensive introduction to the subject. PIERCE BUTLER, *The Origin of Printing in Europe* (1940), a standard work on the subject; HERBERT BUTTERFIELD, *The Origins of Modern Science, 1300–1800* (1949), the best non-technical introduction; CARLO M. CIPOLLA, *Guns, Sails and Empires: Technological Innovation and the Early Phases of European Expansion, 1400–1700* (1966), explores the technological origins of Western supremacy; J.H. PARRY, *The Age of Reconaissance,* 2nd ed. (1966), full of information on navigation, shipbuilding, and exploration. PAUL O. KRISTELLER, *Renaissance Thought and Its Sources* (1979), is a collection of essays discussing Renaissance philosophy, theology, science, and literature.

History, 1648–1815: Two multivolume series cover this period in its entirety: *The Rise of Modern Europe,* ed. by W.L. LANGER, 20 vol. (1936–71); and *The New Cambridge Modern History,* planned by SIR GEORGE N. CLARK, 14 vol. (1957–). Each set deals at length with diplomatic and military developments. The annotated bibliographies in the Langer series are useful and valuable. For special studies of war and diplomacy, see GASTON ZELLER, *Les Temps modernes,* vol. 2, *De Louis XIV à 1789* (1955); and ANDRE FUGIER, *La Révolution française et l'Empire Napoléonien* (1954), both diplomatic and military histories at their broadest, linking them to economic organization and social forces. JOHN U. NEF, *War and Human Progress* (1950); and SIR GEORGE N. CLARK, *War and Society in the Seventeenth Century* (1958), convincingly refute the traditional thesis that war and militarism were the matrix of cultural progress. The later 17th century is treated in two admirable studies by JOHN B. WOLF: *The Emergence of the Great Powers, 1685–1715* (1951) and *Louis XIV* (1968). Both are broadly conceived, rich in details, and persuasively written. Also balanced and judicious is PIERRE GOUBERT, *Louis XIV et vingt millions de français* (Eng. trans., *Louis XIV and Twenty Million Frenchmen,* 1970), which makes a strong case for the greatness of Louis XIV. For Marlborough there is WINSTON CHURCHILL's six-volume authoritative tribute to his ancestor, *Marlborough: His Life and Times* (1933–38). The old study by ALFRED ARNETH, *Prinz Eugen von Savoyen,* new ed., 3 vol. (1864), is still basic. Three consecutive volumes in the Langer series deal with the period from the 18th century to the French Revolution: PENFIELD ROBERTS, *The Quest for Security, 1715–1740* (1947); WALTER L. DORN, *Competition for Empire, 1740–1763* (1940); and LEO GERSHOY, *From Despotism to Revolution, 1763–1789* (1944)—all have chapters on the central themes of balance of power diplomacy, competitive militarism, and the global war for markets and colonial goods. Prussia's place in the wars is discussed acutely in PIERRE GAXOTTE, *Frédéric II* (1938; Eng. trans., 1942). V.O. KLIUCHEVSKY, *A History of Russia,* vol. 5 (Eng. trans., 1931), deals with the diplomacy and wars of Catherine II. For orderly introductions to the troubled years from 1789 to 1815, there are LEO GERSHOY, *The French Revolution and Napoleon,* rev. ed. (1964); and GEOFFREY BRUUN, *Europe and the French Imperium, 1799–1814* (1938). Commercial war is discussed with fresh acumen by ELI F. HECKSCHER, in *The Continental System: An Economic Interpretation* (1922, reprinted 1964). There are many reprints of the once authoritative *The Influence of Sea Power upon the French Revolution and Empire, 1793–1812,* by ALFRED T. MAHAN, 2 vol. (1892). ROBERT B. MOWAT, *The Diplomacy of Napoleon* (1924, reprinted 1971), is a useful brief introduction; while EDOUARD DRIAULT, *Napoléon et l'Europe,* 5 vol. (1910–27), is a storehouse of details. Still quite useful and authoritative is THEODORE A. DODGE, *Napoleon: A History of the Art of War,* 4 vol. (1904–07). The settlement at the Congress of Vienna is studied in CHARLES K. WEBSTER, *The Congress of Vienna, 1814–1815* (1919, reprinted 1963); and, in a pragmatic temper, in HENRY A. KISSINGER, *A World Restored: Metternich, Castlereagh and the Problems of Peace, 1812–1822* (1957). GEOFFREY BEST, *War and Society in Revolutionary Europe: 1770–1870* (1982), studies the social effects of frequent military activity.

Enlightenment: NORMAN HAMPSON, *The Enlightenment,* vol. 4 of the *Pelican History of European Thought* (1968), an introduction to the 18th century—full of knowledge and spirit; PRESERVED SMITH, *A History of Modern Culture,* 2 vol. (1930–34), wisdom expressed in good style is not soon out of date; PETER GAY, *The Enlightenment,* vol. 1, *The Rise of Modern Paganism* (1966), vol. 2, *The Science of Freedom* (1969), a work of distinguished scholarly labour mindful of the complexities in the century's struggle to repudiate Christianity and to rise above classical thought to modernity—the critical bibliographical essays are indispensable for any serious student; ERNST CASSIRER, *Die Philosophie der Aufklärung* (1932; Eng. trans., *The Philosophy of the Enlightenment,* 1951), a classic work by a profound German scholar; PAUL HAZARD, *La Crise de la conscience européenne* (1935; Eng. trans., *The European Mind, 1680–1715,* 1953) and *La Pensée européenne au XVIIIéme siècle, de Montesquieu à Lessing* (1946; Eng. trans., *European Thought in the Eighteenth Century, from Montesquieu to Lessing,* 1954), by a contrasting French scholar who puts a light touch on learning; CARL LOTUS BECKER, *The Heavenly City of the Eighteenth Century Philosophers* (1932; paperback reprint, 1959), a much appreciated set of lectures in which a gentle cynic finds it amusing that the philosophers were doing essentially what their forebears had been doing (PETER GAY has reproved him in *The Party of Humanity: Essays in the French Enlightenment,* 1964); GEORGE R. HAVENS, *The Age of Ideas* (1955), scholarly but perhaps over-enthusiastic biographies of the principal French *philosophes;* GEORGE H. SABINE, *A History of Political Theory,* 3rd ed. (1961), the best survey of this vast field by a philosopher of great and dispassionate learning; A.R. HALL, *The Scientific Revolution, 1500–1800,* 2nd ed. (1962), a standard work on an essential aspect of the Enlightenment; J.B. BURY, *The Idea of Progress: An Inquiry into Its Origin and Growth* (1920), the famous classic, which might profitably be read with CHARLES FRANKEL, *The Faith of Reason: The Idea of Progress in the French Enlightenment* (1948); and HENRY VYVERBERG, *Historical Pessimism in the French Enlightenment* (1958). MARGARET C. JACOB, *The Radical Enlightenment* (1981), is a study of how the more radical thinkers of the Enlightenment affected subsequent history.

History, 1815–1914: The best comprehensive treatment of the subject is contained in the two volumes by PIERRE RENOUVIN: *Le XIX^e Siècle:* vol. 1, *De 1815 à 1871* (1954) and vol. 2, *De 1871 à 1914* (1955). A.J.P. TAYLOR, *The Struggle for Mastery in Europe, 1848–1918* (1954), covers most of the period in the author's provocative, if unorthodox, fashion. RENE ALBRECHT-CARRIE, *A Diplomatic History of Europe Since the Congress of Vienna,* 2nd ed. (1973), which covers the 20th century as well, is a dependable, balanced treatment; it can be supplemented by the same author's *Concert of Europe* (1968), a critical documentary collection that deals with the major episodes and settlements of 19th-century international affairs of Europe. ROBERT B. MOWAT, *A History of European Diplomacy, 1815–1914* (1922), is still useful though dated. Two works dealing with British policy, both of them of high quality, deserve special mention: ADOLPHUS WARD and GEORGE P. GOOCH (eds.), *The Cambridge History of British Foreign Policy, 1783–1919,* 3 vol. (1922–23, reprinted 1970); and ROBERT W. SETON-WATSON, *Britain in Europe, 1789–1914* (1937, reprinted 1968). The second half of the first, and the second and third volumes of a collective Russian work, published under the editorship of VLADIMIR P. POTEMKIN and available in French translation, *Histoire de la diplomatie* (1946–47), covers the 19th century; it has the advantages and the limitations that can be expected of a Soviet treatise in this domain. Two books with a revisionist approach are the pioneering work of SIDNEY B. FAY, *The Origins of the World War,* 2nd ed. rev., 2 vol. (1966); and of ERICH BRANDENBURG, *Von Bismarck zum Weltkriege* (1924; Eng. trans., *From Bismarck to the World War,* 1927). The two masterful volumes by WILLIAM L. LANGER, *European Alliances and Alignments, 1871–1890,* 2nd ed. (1950), and *The Diplomacy of Imperialism, 1890–1902,* 2nd ed. (1951), which are also somewhat informed by the revisionist tendency, are very detailed and dependable. Representing a different orientation, but of similarly high quality, is BERNADOTTE E. SCHMITT, *Triple Alliance and Triple Entente* (1934, reprinted 1971). PIERRE RENOUVIN, *Les Origines immédiates de la guerre, 28 juin–4 août, 1914,* 2nd ed. rev. (1927; Eng. trans., *The Immediate Origins of the War, 28 June–4 August 1914,* 1928) deals exclusively with the July 1914 crisis. The single best and most balanced work is LUIGI ALBERTINI, *Le origini della guerra del 1914,* 3 vol. (1942–43; Eng. trans., *The Origins of the War of 1914,* 3 vol., 1952–57), which has the merit of having had at its disposal a fuller documentation. The last two volumes of it, like the second of Fay's work, deal with the July crisis. V.G. KIERNAN, *From Conquest to Collapse: European Empires from 1815 to 1960* (1982), is a thorough investigation of military aspects of colonial rule.

European culture since 1800: EUGEN WEBER, *Paths to the Present* (1960), is a full and well-organized collection of documents and extracts covering the period from Romanticism to Existentialism. For the corresponding historical narrative, the *Columbia History of the World,* ch. 66–101 (1972), together with E.J. HOBSBAWM, *The Age of Revolution, 1789–1848* (1962), is useful. HAROLD T. PARKER, *The Cult of Antiquity and the French Revolutionaries* (1937, reprinted 1965), is an excellent introduction; as is CRANE BRINTON, *The Political Ideas of the English Romanticists,* rev. ed. (1962). JACQUES BARZUN, *Classic, Romantic, and Modern,* rev. ed. (1962), fills out the interpretation given in the present article. ROBERT C. BINKLEY, *Realism and Nationalism: 1852–1871* (1935); and C.J.H. HAYES, *A Generation of Materialism: 1871–1900* (1941), amplify the political and economic narrative and also interpret culture. EGON FRIEDELL, *Kulturgeschichte der Neuzeit: Die Krisis der europäischen Seele von der schwarzen Pest bis zum Weltkrieg,* 3 vol. (1927–31; Eng. trans., *A Cultural History of the Modern Age,* 3 vol., 1930–32, reprinted 1952–54), is uneven in pace but full of brilliant insights.

History since 1914: The causes of World War I, from 1878, are best presented by LUIGI ALBERTINI, *Le origini della guerra del 1914,* 3 vol. (1942–43; Eng. trans., *The Origins of the War of 1914,* 3 vol., 1952–57). More controversial is FRITZ FISCHER, *Griff nach der Weltmacht,* 3rd ed. On the peacemaking after World War I, the standard narrative in English remains that of H.W.V. TEMPERLEY (ed.), *A History of the Peace Conference of Paris,* 6 vol. (1920–24); which, however, is supplemented by ARNO J. MAYER, *Politics and Diplomacy of Peacemaking: Containment and Counter-revolution at Versailles, 1918–1919* (1967). For the 1920s and 1930s, the reader should first consult PIERRE RENOUVIN, *Les Crises du XX^e siècle,* 2 vol. (1957–58; Eng. trans., *War and Aftermath, 1914–1929* and *World War II and Its Origins: International Relations 1929–45,* both 1968). A less dispassionate account is given by WINSTON CHURCHILL in *The Gathering Storm* (1948). ARNOLD WOLFERS, *Britain and France Between Two Wars: Conflicting Strategies of Peace Since Versailles* (1940), explains the weakness of the West. For World War II there is a good survey for the general reader by BASIL COLLIER, *A Short History of the Second World War* (1967). On the decline of Europe as the historical centre of world politics, see HAJO HOLBORN, *The Political Collapse of Europe* (1959), and JOHN LUKACS, *The Last European War: September 1939–December 1941* (1976). ROLAND N. STROMBERG, *Redemption by War: The Intellectuals and 1914* (1982), studies the widespread support given war efforts by European intellectuals.

The History of European Overseas Exploration and Empires

The motives that spur human beings to examine their environment are many. Strong among them are the satisfaction of curiosity, the pursuit of trade, the spread of religion, and the desire for security and political power. At different times and in different places, different motives are dominant. Sometimes one motive inspires the promoters of discovery, and another motive may inspire the individuals who carry out the search. Still other motives draw settlers to the new territory.

Since the settlement of the European continent, its people have shown an inclination to explore and expand from their geographic centre. Exploration of the Mediterranean world led to contacts with northern and western Europeans that extended the knowledge and culture of both groups. The impetus toward expansion, demonstrated in the establishment of the Hellenistic and Roman empires, continued as Christianity spread through Europe and beyond. Colonization of conquered territories was undertaken, especially by the Romans, but on a much smaller scale than that which followed in the early modern world. The major benefit of early, indeed of all, European expansion was the cultural enrichment that resulted from contact with other civilizations. (J.B.Mi./Ed.)

The major period of European colonization had its origin with the Renaissance, the development of modern science, and the great voyages of discovery. This period began about 1500 and reached its peak in the early 1900s, when the last independent territories of Asia and Africa were parcelled out. Following World War II the strengthening of nationalistic movements opposed to colonialism and the erosion of dominance caused by the modernization of economic systems brought about the decline of the colonial empires.

For a discussion of the society that engaged in these explorations, and their effects on intra-European affairs, see EUROPEAN HISTORY AND CULTURE. The earliest European empires are discussed in GREEK AND ROMAN CIVILIZATIONS, ANCIENT.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 961, and the *Index.*

This article is divided into the following sections:

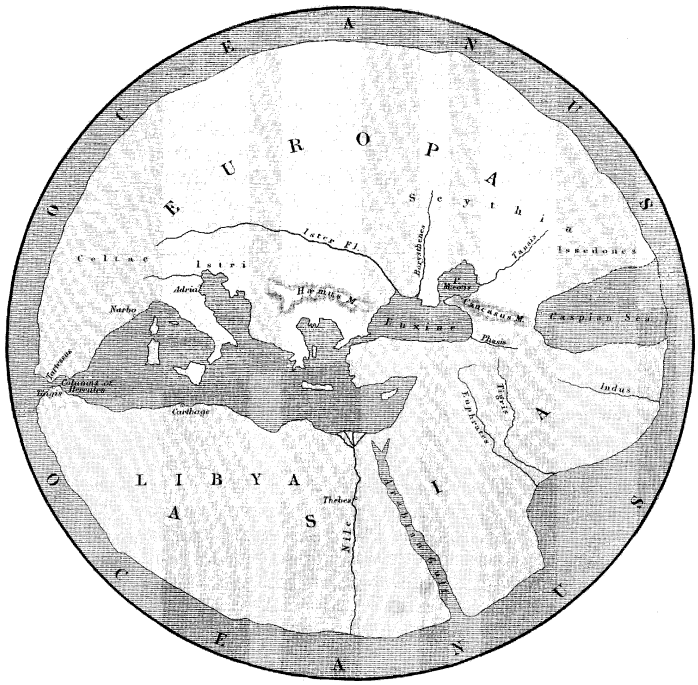
EUROPEAN EXPLORATION

The threads of geographical exploration are continuous and, being entwined one with another, are difficult to separate; three major phases of investigation may nevertheless be distinguished. The first phase is the exploration of the Old World centred on the Mediterranean Sea; the second is the so-called Age of Discovery, during which, in the search for sea routes to Cathay (the name by which China was known to medieval Europe), a New World was found; the third is the establishment of the political, social, and commercial relationships of the New World to the Old and the elucidation of the major physical features of the continental interiors—in short, the delineation of the modern world.

The exploration of the Old World

From the time of the earliest recorded history to the beginning of the 15th century, Western knowledge of the world widened from a river valley surrounded by mountains or desert (the views of Babylonia and Egypt) to a Mediterranean world with hinterlands extending from the Sahara to the Gobi deserts and from the Atlantic to the Indian oceans (the view of Greece and Rome). It later expanded again to include the far northern lands beyond the Baltic and another and dazzling civilization in the Far East (the medieval view).

The earliest known surviving map, dating probably from



Map showing the geographical knowledge of the world derived from the writings of Hecataeus of Miletus, *c.* 500 BC.
By courtesy of the Library of Congress, Washington, D.C.

The earliest known map

the time of Sargon of Akkad (about 2334–2279 BC), shows canals or rivers—perhaps the Tigris and a tributary—and surrounding mountains. The rapid colonization of the shores of the Mediterranean and of the Black Sea by Phoenicia and the Greek city-states in the first millennium BC must have been accompanied by the exploration of their hinterlands by countless unknown soliders and traders. Herodotus prefaces his *History* (written in the 5th century BC) with a geographical description of the then known world: this introductory material reveals that the coastlines of the Mediterranean and the Black Sea had by then been explored.

Stories survive of a few men who are credited with bringing new knowledge from distant journeys. Herodotus tells of five young adventurers of the tribe of the Nasamones living on the desert edge of Cyrenaica in North Africa, who journeyed southwest for many months across the desert, reaching a great river flowing from west to east; this presumably was the Niger, although Herodotus thought it to be the Upper Nile.

EXPLORATION OF THE ATLANTIC COASTLINES

Beyond the Pillars of Hercules (the Strait of Gibraltar), the Carthaginians (from the Phoenician city of Carthage in what is now Tunisia), holding both shores of the strait, early ventured out into the Atlantic. A Greek translation of a Punic (Carthaginian) inscription states that Hanno, a Carthaginian, was sent forth about 500 BC with 60 ships and 30,000 colonists "to found cities." Even allowing for a possible great exaggeration of numbers, this expedition, if it occurred, can hardly have been the first exploratory voyage along the coast of West Africa; indeed, Herodotus reports that Phoenicians circumnavigated the continent about 600 BC. Some scholars think that Hanno reached only the desert edge south of the Atlas; other scholars identify the "deep river infested with crocodiles and hippopotamuses" with the Sénégal River; and still others believe that the island where men "scampered up steep rocks and pelted us with stones" was an island off the coast of Sierra Leone. There is no record that Hanno's voyage was followed up before the era of Henry the Navigator, a Portuguese prince of the 15th century.

About the same time, Himilco, another Carthaginian, set forth on a voyage northward; he explored the coast of Spain, reached Brittany, and in his four-month cruise may have visited Britain. Two centuries later, about 300 BC, Carthaginian power at the gate of the Mediterranean temporarily slackened as a result of squabbles with the Greek city of Syracuse on the island of Sicily, so Pytheas, a Greek explorer of Massilia (Marseille), sailed through. His story is known only from fragments of the work of a contemporary historian, Timaeus (who lived in the 4th and 3rd centuries BC), as retold by the Roman savant Pliny the Elder, the Greek geographer Strabo, and the Greek historian Diodorus Siculus, all of whom were critical of its truth. It is probable that Pytheas, having coasted the shores of the Bay of Biscay, crossed from the island of Ouessant (Ushant), off the French coast of Brittany, to Cornwall in southwestern England, perhaps seeking tin. He may have sailed around Britain; he describes it as a triangle and also relates that the inhabitants "harvest grain crops by cutting off the ears . . . and storing them in covered granges." Around Thule, "the northernmost of the British Isles, six days sail from Britain," there is "neither sea nor air but a mixture like sea-lung . . . binds everything together," a reference perhaps to drift ice or dense sea fog. Thule has been identified with Iceland (too far north), with Mainland island of the Shetland group (too far south), and perhaps, most plausibly, with Norway. Pytheas returned to Brittany and explored "beyond the Rhine"; he may have reached the Elbe. The voyage of Pytheas, like that of Hanno, does not seem to have been followed up. Herodotus concludes by saying, "whether the sea girds Europe round on the north none can tell."

The Northmen

It was not Mediterranean folk but Northmen from Scandinavia, emigrating from their difficult lands centuries later, who carried exploration farther in the North Atlantic. From the 8th to the 11th century bands of Northmen, mainly Swedish, trading southeastward across the Russian plains, were active under the name of Varangians in the ports of the Black Sea. At the same time other groups, mainly Danish, raiding, trading, and settling along the coasts of the North Sea, arrived in the Mediterranean in the guise of Normans. Neither the Swedes nor the Danes travelling in these regions were exploring lands that were unknown to civilized Europeans, but it is doubtless that contact with them brought to these Europeans new knowledge of the distant northern lands.

It was the Norsemen of Norway who were the true explorers though, since little of their exploits was known to contemporaries and that little soon forgotten, they perhaps added less to the common store of Europe's knowledge than their less adventurous compatriots. About AD 890, Ohthere of Norway, "desirous to try how far that country extended north," sailed round the North Cape, along the coast of Lapland to the White Sea. But most Norsemen sailing in high latitudes explored not eastward but westward. Sweeping down the outer edge of Britain, settling in Orkney, Shetland, the Hebrides, and Ireland, they then voyaged on to Iceland, where in 870 they settled among Irish colonists who had preceded them by some two centuries. The Norsemen may well have arrived piloted by Irish sailors; and Irish refugees from Iceland, fleeing before the Norsemen, may have been the first discoverers of Greenland and Newfoundland, although this is mere surmise. The saga of Erik the Red (*Eiríks saga rauda;* also called *Thorfinns saga Karlsefnis*), gives the story of the Norse discovery of Greenland in 982; the west coast was explored, and at least two settlements were established on it. About AD 1000, one Bjarni Herjulfsson, on his way from Iceland to Greenland, was blown off course far to the southwest; he saw an unknown shore and returned to tell his tale. Leif, Erik's son, together with some 30 others, set out in 1001 to explore. They probably reached the coasts of Labrador and Newfoundland; some think that the farthest point south reached by the settlers, as described in the sagas, fits best with Maryland or Virginia, but others contend that the lands about the Gulf of St. Lawrence are more probably designated. The area was named Vinland, as grapes grew there, but it has been suggested that the "grapes" referred to were in fact cranberries. Attempts at colonization were unsuccessful; the Norsemen withdrew; and, although the Greenland colonies lingered on for some four centuries, little knowledge of these first discoveries came down to colour the vision of the seamen of

Cádiz or Bristol; the voyages of Christopher Columbus and John Cabot had their strongest inspirations in quite other traditions.

THE EXPLORATION OF THE COASTLINES OF THE INDIAN OCEAN AND THE CHINA SEA

Trade, across the land bridges and through the gulfs linking those parts of Asia, Africa, and Europe that lie between the Mediterranean and Arabian seas, was actively pursued from very early times. It is therefore not surprising that exploratory voyages early revealed the coastlines of the Indian Ocean. Herodotus wrote of Necho II, king of Egypt in the late 7th and early 6th centuries BC, that "when he stopped digging the canal . . . from the Nile to the Arabian Gulf . . . [he] sent forth Phoenician men in ships ordering them to sail back by the Pillars of Hercules." According to the story, this, in three years, they did. Upon their return, "they told things . . . unbelievable by me," says Herodotus, "namely that in sailing round Libya they had the sun on the right hand." Whatever he thought of the story of the sun, Herodotus was inclined to believe in the voyage: "Libya, that is Africa, shows that it has sea all round except the part that borders on Asia." Strabo records another story with the same theme: one Eudoxus, returning from a voyage to India in about 108 BC, was blown far to the south of Cape Guardafui. Where he landed he found a wooden prow with a horse carved on it, and he was told by the Africans that it came from a wrecked ship of men from the west.

The campaigns of Alexander the Great

About 510 BC, Darius the Great, king of Persia, sent one of his officers, Scylax of Caria, to explore the Indus. Scylax travelled overland to the Kabul River, reached the Indus, followed it to the sea, sailed westward, and, passing by the Persian Gulf (which was already well known), explored the Red Sea, finally arriving at Arsinoë, near modern Suez. The greater part of the campaigns of the famous conqueror Alexander the Great were military exploratory journeys. The earlier expeditions through Babylonia and Persia were through regions already familiar to the Greeks, but later ones through the great tract of land from the south of the Caspian Sea to the mountains of the Hindu Kush brought the Greeks much new geographical knowledge. Alexander and his army crossed the mountains to the Indus Valley and then made a westward march from the lower Indus to Susa through the desolate country along the southern edge of the Iranian plateau; Nearchus, his admiral, in command of the naval forces of the expedition, waited for the favourable monsoon and then sailed from the mouth of the Indus to the mouth of the Euphrates, exploring the northern coast of the Persian Gulf on his way.

As Roman power grew, increasing wealth brought increasing demands for Oriental luxuries; this led to great commercial activity in the eastern seas. As the coasts became well known, the seasonal character of the monsoonal winds was skillfully used; the southwest monsoon was long known as Hippalus, named for a sailor who was credited with being the first to sail with it direct from the Gulf of Aden to the coast of the Indian peninsula. During the reign of the Roman emperor Hadrian in the 1st century BC, Western traders reached Siam (now Thailand), Cambodia (now Kampuchea), Sumatra, and Java; a few also seem to have penetrated northward to the coast of China. In AD 161, according to Chinese records, an "embassy" came from the Roman emperor Marcus Aurelius to the emperor Huan Ti, bearing goods that Huan Ti gratefully received as "tribute." Ptolemy, however, does not know of these voyages: he sweeps his peninsula of Colmorgo (Malay) southwestward to join the eastward trend of his coast of Africa, thus creating a closed Indian Ocean. He presumably did not believe the story of the circumnavigation of Africa. As the 2nd century AD passed, and Roman power declined, trade with the eastern seas did not cease but was gradually taken over by Ethiopians, Parthians, and Arabs. The Arabs, most successful of all, dominated eastern sea routes from the 3rd to the 15th century. In the tales of derring-do of Sindbad the Sailor (a hero of the collection of Arabian tales called *The Thousand and One Nights*), there may be found, behind the fiction, the knowledge of these adventurous Arab sailors and traders, supplying detail to fill in the outline of the geography of the Indian Ocean.

THE LAND ROUTES OF CENTRAL ASIA

The prelude to the Age of Discovery, however, is to be found neither in the Norse explorations in the Atlantic nor in the Arab activities in the Indian Ocean but, rather, in the land journeys of Italian missionaries and merchants that linked the Mediterranean coasts to the China Sea. Cosmas Indicopleustes, an Alexandrian geographer writing in the 6th century, knew that Tzinitza (China) could be reached by sailing eastward, but he added: "one who comes by the overland route from Tzinitza to Persia makes a very short cut." Goods had certainly passed this way since Roman times, but they usually changed hands at many a mart, for disorganized and often warring tribes lived along the routes. In the 13th century the political geography changed. In 1216 a Mongol chief assumed the title of Genghis Khan and, after campaigns in China that gave him control there, turned his conquering armies westward. He and his successors built up an enormous empire until, in the late 13th century, one of them, Kublai Khan, reigned supreme from the Black Sea to the Yellow Sea. Europeans of perspicacity saw the opportunities that friendship with the Mongol power might bring. If Christian Europe could only convert the Mongols, this would at one and the same time heavily tip the scales against Muslim and in favour of Christian power and also give political protection to Christian merchants along the silk routes to the legendary sources of wealth in China. With these opportunities in mind, Pope Innocent IV sent friars to "diligently search out all things that concerned the state of the Tartars," and to exhort them "to give over their bloody slaughter of mankind and to receive the Christian faith." Among others, Giovanni da Pian del Carpini in 1245 and Willem van Ruysbroeck in 1253 went forth to follow these instructions. Travelling the great caravan routes from southern Russia, north of the Caspian and Aral seas and north of the Tien Shan (Tien Mountains), both Carpini and Ruysbroeck eventually reached the court of the emperor at Karakorum. Carpini returned confident that the Emperor was about to become a Christian; Ruysbroeck told of the city in Cathay "having walls of silver and towers of gold"; he had not seen it but had been "credibly informed" of it.

Travels of the Polos

But the greatest of the 13th-century travellers in Asia were the Polos, wealthy merchants of Venice. In 1260 the brothers Nicolo and Maffeo Polo set out on a trading expedition to the Crimea. After two years they were ready to return to Venice, but, finding the way home blocked by war, they travelled eastward to Bukhara (now in the Uzbek Soviet Socialist Republic in the Soviet Union), where they spent another three years. They then accepted an invitation to accompany a party of Tatar envoys returning to the court of Kublai Khan at Cambaluc, near Peking. The Khan received them well, provided them with a gold tablet as a safe-conduct back to Europe, and gave them a letter begging the pope to send "some hundred wise men, learned in the law of Christ and conversant with the seven arts to preach to his people." The Polos arrived home, "having toiled three years on the way," to find that Pope Clement IV was dead. Two years later they set off again, travelling without the wise men but taking with them Nicolo's son, Marco Polo, then a youth of 17. (Marco kept detailed notes of all he saw and, late in life when a captive of the Genoese, dictated to a fellow prisoner a book containing an account of his travels.) This time the Polos took a different route: starting from the port of Hormuz on the Persian Gulf, they crossed Persia to the Pamirs and then followed a caravan route along the southern edge of the Tarim Basin and Gobi Desert to Cambaluc. Knowledge of the route is interesting, but the great contribution of Marco Polo to the geographical knowledge of the West lay in his vivid descriptions of the East. He had great opportunities of seeing China and appreciating its life, for he was taken into the service of the Khan and was sent as an administrator to great cities, busy ports, and remote provinces, with instructions to write full reports. In his book he described how, upon

every main highroad, at a distance apart of 25 or 30 miles (40 to 50 kilometres), there were stations, with houses of accommodation for travellers, with 400 good horses kept in constant readiness at each station. He also reported that, along the roads, the Great Khan had caused trees to be planted, both to provide shade in summer and to mark the route in winter when the ground was covered with snow. Marco Polo lived and worked in western China, visiting the provinces of Shensi, Szechwan, and Yunnan, as well as the borders of Burma. He frequently visited "the noble and magnificent city of Quinsay [Hang-chou], a name that signifies the Celestial City and which it merits from its pre-eminence to all others in the world in point of grandeur and beauty." Cipango (Japan) he did not visit, but he heard about it from merchants and sailors: "it is situated at a distance of 1,500 miles from the mainland.... They have gold in the greatest abundance, its sources being inexhaustible." The most detailed descriptions and the greatest superlatives were reserved for Cambaluc, capital of Cathay, whose splendours were beyond compare; to this city, he said,

> everything that is most rare and valuable in all parts of the world finds its way:... for not fewer than 1,000 carriages and pack-horses loaded with raw silk make their daily entry; and gold tissues and silks of various kinds are manufactured to an immense extent.

No wonder that, when Europe learned of these things, it became enthralled. After 17 years, the Venetians were permitted to depart; they returned to Europe by sea. After visiting Java they sailed through the Strait of Malacca (again proving the error of Ptolemy); and, landing at Hormuz, they travelled cross-country to Armenia, and so home to Venice, which they reached in 1295.

Other European travellers in Asia

A few travellers followed the Polos. Giovanni da Montecorvino, a Franciscan friar from Italy, became archbishop of Peking and lived in China from 1294 to 1328. Friar Oderic of Pordenone, an Italian monk, became a missionary, journeying throughout the greater part of Asia between 1316 and 1330. He reached Peking by way of India and Malaya, then travelled by sea to Canton; he returned to Europe by way of Central Asia, visiting Tibet in 1325—the first European to do so. Friar Oderic's account of his journeys had considerable influence in his day: it was from it that the spurious traveller, the English writer Sir John Mandeville, quarried most of his stories.

Ibn Baṭṭūṭah, an Arab of Tangier, journeyed farther perhaps than any other medieval traveller. In 1325 he set out to make the traditional pilgrimage to Mecca, and in some 30 years he visited the greater part of the Old World, covering, it has been said, more than 75,000 miles. He was the first to explore much of Arabia; he travelled extensively in India; he reached Java and Southeast Asia. Then toward the end of his life he returned to the west, where, after visiting Spain, he explored western Sudan "to the northernmost province of the Negroes." He reached the Niger, which he called the Nile, and was astonished by the huge hippopotamuses "taking them to be elephants." When he finally returned to Fès in Morocco he "kissed the hand of the Commander of the Faithful the Sultan... and settled down under the wing of his bounty." He wrote a vivid and perspicacious account of his travels, but his book did not become known to Christian Europe for centuries. It was Marco Polo's book that was the most popular of all. Some 138 manuscripts of it survive: it was translated before 1500 into Latin, German, and Spanish, and the first English translation was published in 1577. For centuries Europe's maps of the Far East were based on the information provided by Marco Polo; even as late as 1533 Johannes Schöner, the German maker of globes, wrote:

> Behind the Sinae and the Ceres [legendary cities of Central Asia]... many countries were discovered by one Marco Polo... and the sea coasts of these countries have now recently again been explored by Columbus and Amerigo Vespucci in navigating the Indian Ocean.

Columbus possessed and annotated a copy of the Latin edition (1483–85) of Marco Polo's book, and in his journal he identified many of his own discoveries with places that Marco Polo describes.

Thus, with Ptolemy in one hand and Marco Polo in the other, the European explorers of the Age of Discovery set forth to try to reach Cathay and Cipango by new ways; Ptolemy promised that the way was short; Marco Polo promised that the reward was great.

The Age of Discovery

In the 100 years from the mid-15th to the mid-16th century, a combination of circumstances stimulated men to seek new routes; and it was new routes rather than new lands that filled the minds of kings and commoners, scholars and seamen. First, toward the end of the 14th century, the vast empire of the Mongols was breaking up; thus, Western merchants could no longer be ensured of safe-conduct along the land routes. Second, the growing power of the Ottoman Turks, who were hostile to Christians, blocked yet more firmly the outlets to the Mediterranean of the ancient sea routes from the East. Third, new nations on the Atlantic shores of Europe were now ready to seek overseas trade and adventure.

THE SEA ROUTE EAST BY SOUTH TO CATHAY

Henry the Navigator

Henry the Navigator, prince of Portugal, initiated the first great enterprise of the Age of Discovery—the search for a sea route east by south to Cathay. His motives were mixed. He was curious about the world; he was interested in new navigational aids and better ship design and was eager to test them; he was also a crusader and hoped that, by sailing south and then east along the coast of Africa, Arab power in North Africa could be attacked from the rear. The promotion of profitable trade was yet another motive; he aimed to divert the Guinea trade in gold and ivory away from its routes across the Sahara to the Moors of Barbary (North Africa) and instead channel it via the sea route to Portugal.

Expedition after expedition was sent forth throughout the 15th century to explore the coast of Africa. In 1445 the Portuguese navigator Dinís Dias reached the mouth of the Sénégal, which "men say comes from the Nile, being one of the most glorious rivers of Earth, flowing from the Garden of Eden and the earthly paradise." Once the desert coast had been passed, the sailors pushed on: in 1455 and 1456 Alvise Ca' da Mosto made voyages to Gambia and the Cape Verde Islands. Prince Henry died in 1460 after a career that had brought the colonization of the Madeira Islands and the Azores and the traversal of the African coast to Sierra Leone. Henry's captain, Diogo Cão, discovered the Congo River in 1482. All seemed promising; trade was good with the riverine peoples, and the coast was trending hopefully eastward. Then the disappointing fact was realized: the head of a great gulf had been reached, and, beyond, the coast seemed to stretch endlessly southward. Yet, when Columbus sought backing for his plan to sail westward across the Atlantic to the Indies, he was refused—"seeing that King John II [of Portugal] ordered the coast of Africa to be explored with the intention of going by that route to India."

The search for routes to Asia

King John II sought to establish two routes: the first, a land and sea route through Egypt and Ethiopia to the Red Sea and the Indian Ocean and, the second, a sea route around the southern shores of Africa, the latter an act of faith, since Ptolemy's map showed a landlocked Indian Ocean. In 1487, a Portuguese emissary, Pêro da Covilhã, successfully followed the first route; but, on returning to Cairo, he reported that, in order to travel to India, the Portuguese "could navigate by their coasts and the seas of Guinea." In the same year another Portuguese navigator, Bartolomeu Dias, found encouraging evidence that this was so. In 1487 he rounded the Cape of Storms in such bad weather that he did not see it, but he satisfied himself that the coast was now trending northeastward; before turning back, he reached the Great Fish River, in what is now South Africa. On the return voyage, he sighted the Cape and set up a pillar upon it to mark its discovery.

The seaway was now open, but eight years were to elapse before it was exploited. In 1492 Columbus had apparently reached the East by a much easier route. By the end of the decade, however, doubts of the validity of Columbus' claim were current. Interest was therefore renewed

World map by J.M. Contarini, 1506, depicting the expanding horizons becoming known to European geographers in the Age of Discovery.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

in establishing the sea route south by east to the known riches of India. In 1497 a Portuguese captain, Vasco da Gama, sailed in command of a fleet under instructions to reach Calicut, on India's west coast. This he did after a magnificent voyage around the Cape of Storms (which he renamed the Cape of Good Hope) and along the unknown coast of East Africa. Yet another Portuguese fleet set out in 1500, this one being under the command of Pedro Álvarez Cabral; on the advice of da Gama, Cabral steered southwestward to avoid the calms of the Guinea coast; thus, en route for Calicut, Brazil was discovered. Soon trading depots, known as factories, were built along the African coast, at the strategic entrances to the Red Sea and the Persian Gulf, and along the shores of the Indian peninsula. In 1511 the Portuguese established a base at Malacca (now Melaka, Malaysia), commanding the straits into the China Sea; in 1511 and 1512, the Moluccas, or Spice Islands, and Java were reached; in 1557 the trading port of Macau was founded at the mouth of the Canton River. Europe had arrived in the East. It was in the end the Portuguese, not the Turks, who destroyed the commercial supremacy of the Italian cities, which had been based on a monopoly of Europe's trade with the East by land. But Portugal was soon overextended; it was therefore the Dutch, the English, and the French who in the long run reaped the harvest of Portuguese enterprise.

Some idea of the knowledge that these trading explorers brought to the common store may be gained by a study of contemporary maps. The map of the German Henricus Martellus, published in 1492, shows the shores of North Africa and of the Gulf of Guinea more or less correctly and was probably taken from numerous seamen's charts. The delineation of the west coast of southern Africa from the Guinea Gulf to the Cape suggests a knowledge of the charts of the expedition of Bartolomeu Dias. The coastlines of the Indian Ocean are largely Ptolemaic with two exceptions: first, the Indian Ocean is no longer landlocked; and second, the Malay Peninsula is shown twice—once according to Ptolemy and once again, presumably, according to Marco Polo. The Contarini map of 1506 shows further advances; the shape of Africa is generally accurate, and there is new knowledge of the Indian Ocean, although it is curiously treated. Peninsular India (on which Cananor and Calicut are named) is shown; although too small, it is, however, recognizable. There is even an indication to the east of it of the Bay of Bengal, with a great river running into it. Eastward of this is Ptolemy's India, with the huge island of Taprobane—a muddled representation of the Indian peninsula and Ceylon (now Sri Lanka). East again, as on the map of Henricus Martellus, the Malay Peninsula appears twice. Ptolemy's bonds were hard to break.

THE SEA ROUTE WEST TO CATHAY

It is not known when the idea originated of sailing westward in order to reach Cathay. Many sailors set forth searching for islands in the west; and it was a commonplace among scientists that the east could be reached by sailing west, but to believe this a practicable voyage was an entirely different matter. Christopher Columbus, a Genoese who had settled in Lisbon about 1476, argued that Cipango lay a mere 2,500 nautical miles west of the Canary Islands in the eastern Atlantic. He took 45 instead of 60 nautical miles as the value of a degree; he accepted Ptolemy's exaggerated west–east extent of Asia and then added to it the lands described by Marco Polo, thus reducing the true distance between the Canaries and Cipango by about one-third. He could not convince the Portuguese scientists nor the merchants of Lisbon that his idea was worth backing; but eventually he obtained the support of King Ferdinand and Queen Isabella of Spain.

The sovereigns probably argued that the cost of equipping the expedition would not be very great; the loss, if it failed, could be borne; the gain, should it succeed, was incalculable—indeed, it might divert to Spain all the wealth of Asia.

The voyages of Columbus

On August 3, 1492, Columbus sailed from Palos, Spain, with three small ships manned by Spaniards. From the Canaries he sailed westward, for, on the evidence of the globes and maps in which he had faith, Japan was on the same latitude. If Japan should be missed, Columbus thought that the route adopted would land him, only a little further on, on the coast of China itself. Fair winds favoured him, the sea was calm, and, on October 12, landfall was made on the Bahama island of Guanahaní, which he renamed San Salvador (also called Watling Island, though Samana Cay and other islands have been identified as Guanahaní). With the help of the local Indians, the ships reached Cuba and then Haiti. Although there was no sign of the wealth of the lands of Kublai Khan, Columbus nevertheless seemed convinced that he had reached China, since, according to his reckoning, he was beyond Japan. A second voyage in 1493 and 1494, searching fruitlessly for the court of Kublai Khan, further explored the islands of "the Indies." Doubts seem to have arisen among the would-be colonists as to the identity of the islands since Columbus demanded that all take an oath that Cuba was the southeast promontory of Asia—the Golden Chersonese. On his third voyage, in 1498, Columbus sighted Trinidad, entered the Gulf of Paria, on the coast of what is now Venezuela, and annexed for Spain "a very great continent . . . until today unknown." On a fourth voyage, from 1502 to 1504, he explored the coast of Central America from Honduras to Darien on the Isthmus of Panama, seeking a navigable passage to the west. What passage he had in mind is obscure; if at this point he still believed he had reached Asia, it is conceivable that he sought a way through Ptolemy's Golden Chersonese into the Indian Ocean.

Columbus' tenacity, courage, and skill in navigation make him stand out among the few explorers who have changed substantially ideas about the world. At the time, however, his efforts must have seemed ill-rewarded: he found no emperor's court rich in spices, silks, gold, or precious stones but had to contend with mutinous sailors, dissident colonists, and disappointed sovereigns. He died at Valladolid in 1506. Did he believe to the end that he indeed had reached Cathay, or did he, however dimly, perceive that he had found a New World?

Whatever Columbus thought, it was clear to others that there was much to be investigated, and probably much to be gained, by exploration westward. Not only in Lisbon and Cádiz but also in other Atlantic ports, groups of men congregated in hopes of joining in the search. In England, Bristol, with its western outlook and Icelandic trade, was the port best placed to nurture adventurous seamen. In the latter part of the 15th century, John Cabot, with his wife and three sons, came to Bristol from Genoa or Venice. His project to sail west gained support, and with one small ship, the "Matthew," he set out in May 1497, taking a course due west from Dursey Head, Ireland. His landfall on the other side of the ocean was probably on the northern peninsula of what is now known as Newfoundland. From there, Cabot explored southward, perhaps encouraged to do so, even if seeking a westward passage, by ice in the Strait of Belle Isle. Little is known of John Cabot's first voyage, and almost nothing of his second, in 1498, from which he did not return, but his voyages in high latitudes represented almost as great a navigational feat as those of Columbus.

The coasts between the landfalls of Columbus and of John Cabot were charted in the first quarter of the 16th century by Italian, French, Spanish, and Portuguese sailors. Sebastian Cabot, son of John, gained a great reputation as a navigator and promoter of Atlantic exploration, but whether this was based primarily on his own experience or on the achievements of his father is uncertain. In 1499 Amerigo Vespucci, an Italian merchant living in Seville, together with the Spanish explorer Alonso de Ojeda, explored the north coast of South America from Suriname to the Golfo de Venezuela. His lively and embellished description of these lands became popular, and Waldseemüller, on his map of 1507, gave the name America to the southern part of the continent.

Contarini's map

The 1506 map of Contarini represented a brave attempt to collate the mass of new information, true and false, that accrued from these western voyages. The land explored by Columbus on his third voyage and by Vespucci and de Ojeda in 1499 is shown at the bottom left of the map as a promontory of a great northern bulge of a continent extending far to the south. The northeast coast of Asia at the top left is pulled out into a great peninsula on which is shown a big river and some mountains representing Contarini's concept of Newfoundland and the lands found by the Cabots and others. In the wide sea that separates these northern lands from South America, the West Indies are shown. Halfway between the Indies and the coast of Asia, Japan is drawn. A legend placed between Japan and China reveals the state of opinion among at least some contemporary geographers; it presumably refers to the fourth voyage of Columbus in 1502 and may be an addition to the map. It runs:

> Christopher Columbus, Viceroy of Spain, sailing westwards, reached the Spanish islands after many hardships and dangers. Weighing anchor thence he sailed to the province called Ciambra [a province which then adjoined Cochinchina].

Others did not agree with Contarini's interpretation. To more and more people it was becoming plain that a New World had been found, although for a long time there was little inclination to explore it but instead a great determination to find a way past it to the wealth of Asia. The voyage of the Portuguese navigator Ferdinand Magellan, from 1519 to 1521, dispelled two long-cherished illusions: first, that there was an easy way through the barrier and, second, that, once the barrier was passed, Cathay was near at hand.

Magellan's voyages

Ferdinand Magellan had served in the East Indies as a young man. Familiar with the long sea route to Asia eastward from Europe via the Cape of Good Hope, he was convinced that there must be an easier sea route westward. His plan was in accord with Spanish hopes; five Spanish ships were fitted out in Seville, and in August 1519 they sailed under his command first to the Cape Verde Islands and thence to Brazil. Standing offshore, they then sailed southward along the east coast of South America; the estuary of the Río de la Plata was explored in the vain hope that it might prove to be a strait leading to the Pacific. Magellan's ships then sailed south along the coast of Patagonia. The Gulf of St. George, and doubtless many more small embayments, raised hopes that a strait had been found, only to dash them; at last at Port Julian, at 49°15′ S, winter quarters were established. In September 1520 a southward course was set once more, until, finally, on October 21, Magellan found a strait leading westward. It proved to be an extremely difficult one: it was long, deep, tortuous, rock-walled, and bedevilled by icy squalls and dense fogs. It was a miracle that three of the five ships got through its 325-mile length. After 38 days, they sailed out into the open ocean. Once away from land, the ocean seemed calm enough; Magellan consequently named it the Pacific. The Pacific, however, proved to be of vast extent, and for 14 weeks the little ships sailed on a northwesterly course without encountering land. Short of food and water, the sailors ate sawdust mixed with ship's biscuits and chewed the leather parts of their gear to keep themselves alive. At last, on March 6, 1521, exhausted and scurvy-ridden, they landed at the island of Guam. Ten days later they reached the Philippines, where Magellan was killed in a local quarrel. The survivors, in two ships, sailed on to the Moluccas; thus, sailing westward, they arrived at last in territory already known to the Portuguese sailing eastward. One ship attempted, but failed, to return across the Pacific. The remaining ship, the "Vittoria," laden with spices, under the command of the Spanish navigator Juan Sebastián de Elcano, sailed alone across the Indian Ocean, rounded the Cape of Good Hope, and arrived at Seville on September 9, 1522, with a crew of four Indians and only 17 survivors of the 239 Europeans who had set sail with the expedition three years earlier. Elcano, not having

allowed for the fact that his circumnavigation had caused him to lose a day, was greatly puzzled to find that his carefully kept log was one day out; he was, however, delighted to discover that the cargo that he had brought back more than paid for the expenses of the voyage.

It is fitting to consider this first circumnavigation as marking the close of the Age of Discovery. Magellan and his men had demonstrated that Columbus had discovered a New World and not the route to China and that Columbus' "Indies"—the West Indies—were separated from the East Indies by a vast ocean.

Not all the major problems of world geography were, however, now solved. Two great questions still remained unanswered. Were there "northern passages" between the Atlantic and Pacific oceans more easily navigable than the dangerous Strait of Magellan to the south? Was there a great landmass somewhere in the vastness of the southern oceans—a Terra Australis ("southern land") that would balance the northern continents?

The emergence of the modern world

The centuries that have elapsed since the Age of Discovery have seen the end of dreams of easy routes to the East by the north, the discovery of Australasia and Antarctica in place of Terra Australis Incognita, and the identification of the major features of the continental interiors.

While, as in earlier centuries, traders and missionaries often proved themselves also to be intrepid explorers, in this period of geographical discovery the seeker after knowledge for its own sake played a greater part than ever before.

THE NORTHERN PASSAGES

Roger Barlow, in his *Briefe Summe of Geographie,* written in 1540–41, asserted that "the shortest route, the northern, has been reserved by Divine Providence for England."

The search for a Northeast Passage

The concept of a Northeast Passage was at first favoured by the English: it was thought that, although its entry was in high latitudes, it "turning itself, trendeth towards the southeast . . . and stretcheth directly to Cathay." It was also argued that the cold lands bordering this route would provide a much needed market for English cloth. In 1553 a trading company, later known as the Muscovy Company, was formed with Sebastian Cabot as its governor. Under its auspices numerous expeditions were sent out. In 1553 an expedition set sail under the command of Sir Hugh Willoughby; Willoughby's ship was lost, but the exploration continued under the leadership of its pilot general, Richard Chancellor. Chancellor and his men wintered in the White Sea, and next spring "after much adoe at last came to Mosco." Between 1557 and 1560, another English voyager, Anthony Jenkinson, following up this opening, travelled from the White Sea to Moscow, then to the Caspian, and so on to Bukhara, thus reaching the old east–west trade routes by a new way. Soon, attempts to find a passage to Cathay were replaced by efforts to divert the trade of the ancient silk routes from their traditional outlets on the Black Sea to new northern outlets on the White Sea.

The Dutch next took up the search for the passage. The Dutch navigator William Barents made three expeditions between 1594 and 1597 (when he died in Novaya Zemlya, modern Soviet Union). The English navigator Henry Hudson, in the employ of the Dutch, discovered between 1605 and 1607 that ice blocked the way both east and west of Svalbard (Spitsbergen). Between 1725 and 1729 and from 1734 to 1743, a series of expeditions inspired by the Danish-Russian explorer Vitus Bering attempted the passage from the eastern end, but it was not until 1878–79 that Baron Adolf Erik Nordenskiöld, the Finnish-Swedish scientist and explorer, sailed through it.

The Northwest Passage, on the other hand, also had its strong supporters. In 1576 Humphrey Gilbert, the English soldier and navigator, argued that "Mangia [South China], Quinzay [Hang-chou] and the Moluccas are nearer to us by the North West than by the North East," while John Dee in 1577 set out the view that the Strait of Anian, separating America from Asia, led southwest "along the backeside of Newfoundland." In 1534 Jacques Cartier, the French navigator, explored the St. Lawrence estuary. In 1576 the English explorer Sir Martin Frobisher found the bay named after him. Between 1585 and 1587, the English navigator John Davis explored Cumberland Sound and the western shore of Greenland to 73° N; although he met "a mighty block of ice," he reported that "the passage is most probable and the execution easy." In 1610 Henry Hudson sailed through Hudson Strait to Hudson Bay, confident, before he was set adrift by a mutinous crew, that success was at hand. Between 1612 and 1615, three English voyagers—Robert Bylot, Sir Thomas Button, and William Baffin—thoroughly explored the bay, returning convinced that there was no strait out of it leading westward. As in the quest for a Northeast Passage, interest turned from the search for a route leading to the riches of the East to the exploitation of local resources. Englishmen of the Hudson's Bay Company, founded in 1670 to trade in furs, explored the wide hinterlands of the St. Lawrence estuary and Hudson Bay. Further search for the passage itself did not take place until the 19th century: expeditions led by Sir William Parry (1819–25) and Sir John Franklin (1819–45), as well as more than 40 expeditions sent out to search for Franklin and his party, failed to find the passage. It was left to the Norwegian explorer Roald Amundsen to be the first to sail through the passage, which he did in 1903–05.

EASTWARD VOYAGES TO THE PACIFIC

The discovery of Australia

By the end of the 16th century, Portugal in the East held only the ports of Goa and Diu, in India, and Macau, in China. The English dominated the trade of India, and the Dutch that of the East Indies. It was the Dutch, trading on the fringes of the known world, who were the explorers. Victualling their ships at the Cape, they soon learned that, by sailing east for some 3,000 miles (5,000 kilometres) before turning north, they would encounter favourable winds in setting a course toward the Spice Islands (now the Moluccas). Before long, reports were received of landfalls made on an unknown coast; as early as 1618, a Dutch skipper suggested that "this land is a fit point to be made by ships . . . *in order* to get a fixed course for Java." Thereafter, the west coast of Australia was gradually charted: it was identified by some as the coast of the great southern continent shown on Mercator's map and, by others, as the continent of Loach or Beach mentioned by Marco Polo, interpreted as lying to the south of Malacca (Melaka); Polo, however, was probably describing the Malay Peninsula.

In 1642, a farsighted governor general of the Dutch East India Company, Anthony van Diemen, sent out the Dutch navigator Abel Tasman for the immediate purpose of making an exploratory voyage, but with the ultimate aim of developing trade. Sailing first south then east from Mauritius, Tasman landed on the coast of Tasmania, after which he coasted round the island to the south and, sailing east, discovered the South Island of New Zealand; "we trust that this is the mainland coast of the unknown South land," he wrote. He sailed north without finding Cook Strait, and, making a sweeping arc on his voyage back to the Dutch port of Batavia (now Jakarta, Indonesia), he discovered the Tonga and the Fiji Islands. In 1644, on a second voyage, he traced the north coast of Australia from Cape York (which he thought to be a part of New Guinea) to the North West Cape.

WESTWARD VOYAGES TO THE PACIFIC

The earlier European explorers in the Pacific were primarily in search of trade or booty; the later ones were primarily in search of information.

The traders, for the most part Spaniards, established land portages from harbours on the Caribbean to harbours on the west coast of Central and South America; from the Pacific coast ports of the Americas, they then set a course westward to the Philippines. Many of their ships crossed and recrossed the Pacific without making a landfall; many islands were found, named, and lost, only to be found again without recognition, renamed, and perhaps lost yet again. In the days before longitude could be accurately fixed, such uncertainty was not surprising.

Some voyages—for example, those of Álvaro de Mendaña de Neira, the Spanish explorer, in 1567 and 1568; Mendaña and the Portuguese navigator Pedro Fernández de Quirós in 1595; Quirós and another Portuguese explorer, Luis de Torres, in 1606—had, among other motives, the purpose of finding the great southern continent. Quirós was sure that in Espíritu Santo in the New Hebrides he had found his goal; he "took possession of the site on which is to be founded the New Jerusalem." Torres sailed from there to New Guinea and thence to Manila, in the Philippines. In doing so, he coasted the south shore of New Guinea, sailing through Torres Strait, unaware that another continent lay on his left hand.

The English were rivals of the Spaniards in the search for wealth in unknown lands in the Pacific. Two English seamen, Sir Francis Drake and Thomas Cavendish, circumnavigated the world from west to east in 1577 to 1580 and 1586 to 1588, respectively. One of Drake's avowed objects was the search for Terra Australis. Once he was through Magellan's straits, however, strong winds made him turn north—perhaps not reluctantly. He then sailed along the coast of Peru, surprising and plundering Spanish ships laden with gold, silver, precious stones, and pearls. His fortune made, Drake continued northward perhaps in search of the Northwest Passage. He explored the west coast of North America to 48° N. He returned south to winter in New Albion (California); the next summer he sailed on the Spanish route to Manila, then returned home by the Cape.

Despite the fact that he participated in several buccaneering voyages, the English seaman William Dampier, who was active in the late 17th and early 18th centuries, may be regarded as the first to travel mainly to satisfy scientific curiosity. He wrote: "I was well satisfied enough knowing that, the further we went, the more knowledge and experience I should get, which was the main thing I regarded." His book *A New Voyage Round the World,* published in 1697, further popularized the idea of a great southern continent.

In the late 18th century, the final phase of Pacific exploration occurred. The French sent the explorer Louis-Antoine de Bougainville to the Pacific in 1768. He appears to have been more of a skeptic than many of his contemporaries, for, while he agreed "that it is difficult to conceive such a number of low islands and almost drowned lands without a continent near them," at the same time he maintained that "if any considerable land existed hereabouts we could not fail meeting with it." The British, for their part, commissioned John Byron in 1764 and Samuel Wallis and Phillip Carteret in 1766 "to discover unknown lands and to explore the coast of New Albion." For all the navigational skill and personal endurance shown by captains and crews, the rewards of these voyages in increasing geographical knowledge were not great. The courses sailed were in the familiar waters of the southern tropics; none was through the dangerous waters of higher latitudes.

Cook's voyages in the Pacific

Capt. James Cook, the English navigator, in three magnificent voyages at long last succeeded in demolishing the fables about Pacific geography. He was given command of an expedition to observe the transit of the planet Venus at Tahiti on June 3, 1769; with the observation completed, he carried out his instructions to search the area between 40° and 35° S "until you discover it [Terra Australis] or fall in with the eastern side of the land discovered by Tasman and now called New Zealand." He reached New Zealand, circumnavigated both islands, sailed westward, and on April 19, 1770, made landfall on the eastern coast of Australia. He then turned northward, charting carefully, being well aware of the dangers of the Great Barrier Reef. At Cape York, Cook took possession of the whole eastern coast, to which he gave the name New South Wales. He sailed through Torres Strait, recognizing as he did so that New Guinea was an island. When Cook sailed back to England by Batavia and the Cape, the coastline of the fifth continent was almost complete; only in the south did it still remain unknown. In 1798 to 1799, two British navigators, George Bass and Matthew Flinders, circumnavigated Tasmania, and in 1801–03 Flinders charted the coast of the Great Australian Bight and circumnavigated the continent, thereby proving that there was no strait from the bight to the Gulf of Carpentaria.

In a second voyage, from 1772 to 1775, which in many ways was the greatest of the three, Cook searched systematically for the elusive continent that many still believed might exist. The first summer he examined the area to the south of the Indian Ocean; in the second, he searched the ocean between New Zealand and Cape Horn; and, in the third, the ocean between Cape Horn and the Cape of Good Hope. He sailed home convinced that the great South Pacific continent of the map makers was a fable.

With the exploration of the Pacific completed, interest in a Northwest Passage revived. In 1778 Cook proceeded to latitude 65° N, but he found no way through the ice barrier either to east or to west. He then sailed south to Hawaii, where he was killed in a dispute with the islanders.

Terra Australis Incognita had disappeared: there was now no unknown landmass in the southern oceans. It was Matthew Flinders who suggested that the fifth continent should be named Australia—a name that had long associations with the South Seas and that accorded well with the names of the other continents.

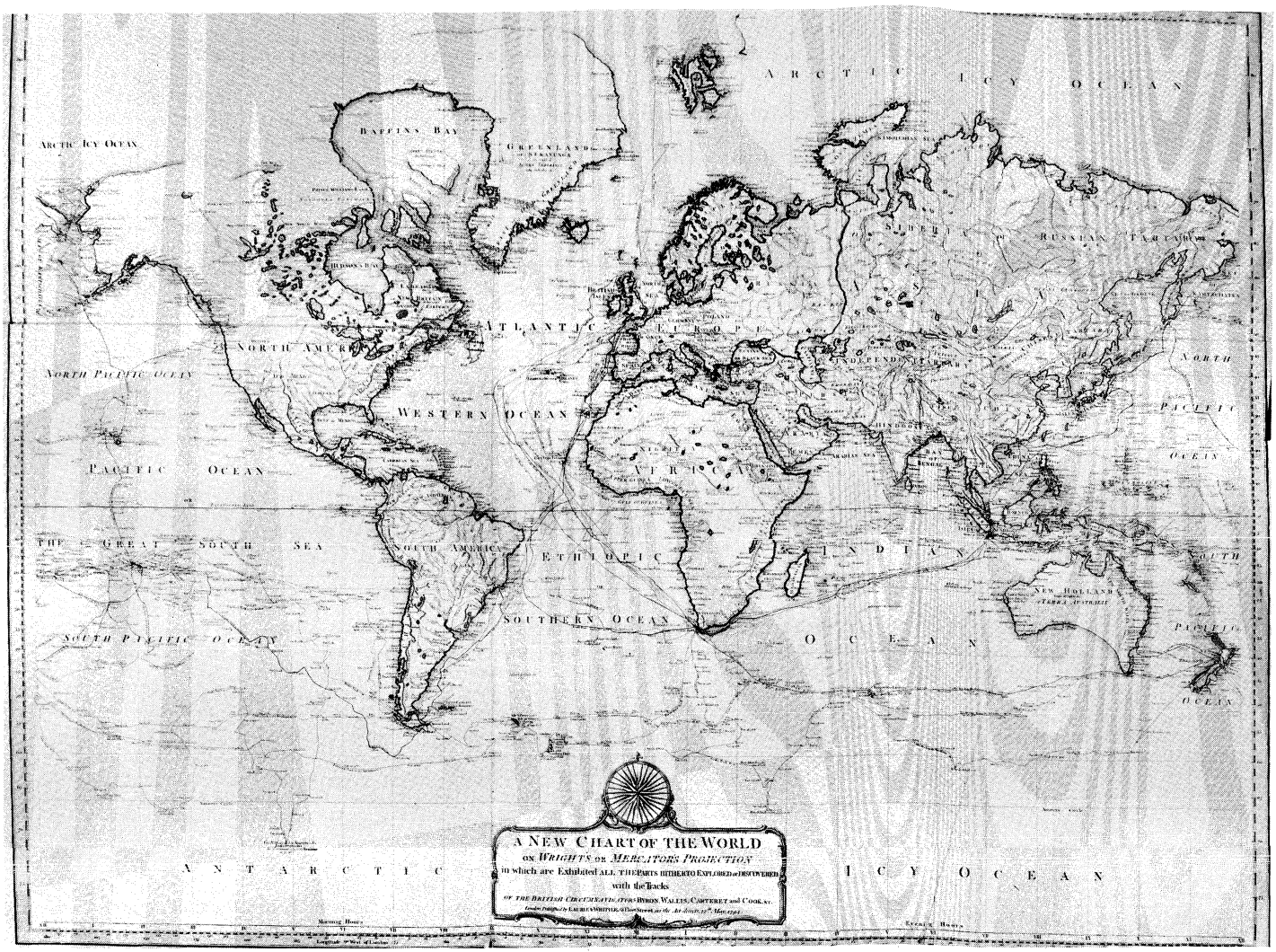
THE CONTINENTAL INTERIORS

At the opening of the 19th century, the major features of Europe, Asia, and North and South America were known; in Africa some classical misconceptions still persisted; inland Australia was still almost blank; and Antarctica was not on the map at all.

Africa. The river systems were the key to African geography. The existence of a great river in the interior of West Africa was known to the Greeks, but in which direction it flowed and whether it found an outlet in the Sénégal, the Gambia, the Congo, or even the Nile were in dispute. A young Scottish surgeon, Mungo Park, was asked to explore it by the African Association of London. In 1796 Park, who had travelled inland from the Gambia, saw "the long sought for majestic Niger flowing slowly *eastwards.*" On a second expedition, attempting to follow its course to the mouth, he was drowned near Bussa, in what is now Nigeria. In 1830 an English explorer, Richard Lander, travelled from the Bight of Benin, on the West African coast, to Bussa, and he then navigated the river down to its mouth, which was revealed as being one of the delta distributaries that, because of the trade in palm oil, were known to traders as "the oil rivers" on the Gulf of Guinea.

Exploration of the Zambezi

The Zambezi, in south central Africa, was not known at all until, in the mid-19th century, the Scottish missionary-explorer David Livingstone crossed the Kalahari from the south, found Lake Ngami, and, hearing of populous areas farther north, came upon the river in midcourse. On a great exploratory journey from 1852 to 1856, the main purpose of which was to expose the slave trade, he first travelled upstream, crossed the watershed between the tributaries of the upper Zambezi and those of the lower Congo, and reached the west coast at Luanda, Angola. From there a year's march brought him back to his starting point near the falls that the Africans called "smoke does sound" but that Livingstone prosaically renamed the Victoria Falls; from here he followed the Zambezi downstream, reaching the east coast at Quelimane, in Portuguese East Africa (Mozambique). On his second journey, sent out by the British government to test the navigability of the lower Zambezi, he explored the Shire (Chire) and Rovuma rivers and reached Lake Nyasa. His last journey, from 1865 to 1871, was undertaken at the behest of the president of Britain's Royal Geographical Society (successor to the African Association) "to solve a question of intense geographical interest . . . namely the watershed or watersheds of southern Africa." On this journey Livingstone investigated the complex drainage system between Lake Nyasa and Lake Tanganyika and explored the headwaters of the Congo. He refused to return to England with the Welsh explorer Henry Morton Stanley, who was sent to his rescue in 1871, because he was still uncertain of the position of the watershed between the Nile and the Congo; he wondered if the Lualaba was perhaps a headstream of the



Thomas Kitchin's "New Chart of the World" (published by Laurie and Whittle, London, 1794), illustrating the state of geographical knowledge before the exploration of Antarctica and some of the continental interiors.

By courtesy of the Royal Geographical Society, London; photograph, John Webb

Nile. He struggled back to the maze of waterways around Lake Bangweulu and died there in 1873.

The source of the Nile

The whereabouts of the source of the Nile had intrigued men since the days of the pharaohs. A Scottish explorer, James Bruce, travelling in Ethiopia in 1770, visited the two fountains in Lake Tana, the source of the Blue Nile, first discovered by the Portuguese priest Paez in 1618. The English explorers Richard Burton and John Speke discovered Lake Tanganyika in 1857. Speke then travelled north alone and reached the southern creek of a lake, which he named Victoria Nyanza. Without exploring farther, he returned to England, sure that he had found the source of the Nile. He was right—but he had not seen the outlet, and Burton did not believe him. In 1862 Speke, travelling with the Scottish explorer James Grant, found the Ripon Falls, in Uganda (now submerged following the construction of a dam for Owen Falls hydroelectric station), and "saw without any doubt that Old Father Nile rises in Victoria Nyanza." Stanley completed the puzzle in 1875; he circumnavigated Victoria Nyanza, crossed to the Lualaba, followed that river to the Congo, and then followed the Congo to its mouth. The pattern made by the river systems of Africa was elucidated at last.

Australia. The interior of Australia also posed a problem: was its heart an inland sea or a desert? This question did not arouse anything approaching the same degree of public interest that was taken in the geography of Africa. Exploration was slow; the early settlers on the east coast found that the valleys led to impassable walls at the valley heads. In 1813 the Australian explorer Gregory Blaxland successfully crossed the Blue Mountains by following a ridge instead of taking a valley route. Rivers were found beyond the mountains, but they did not behave as expected. Another explorer, the Australian John Oxley, in 1818 observed: "on every hill a spring, in every valley a rivulet, but the river itself disappears." He guessed that the great fan of rivers that drained the western slopes of the Great Dividing Range of eastern Australia fell into an inland sea. The Australian Charles Sturt resolved the problem by an imaginative journey made in 1829–30. He embarked on the Murrumbidgee River and was "hurried into a great and noble river [the Murray]." A week later he encountered another big river flowing into the Murray from the north, that he rightly concluded was the Darling, the middle course of which he had explored the year before. The voyage ended when he discovered that the Murray drained into Encounter Bay on the south coast. The heart of Australia was not an inland sea but a vast desert. Many more expeditions were needed to map the continent's major features, but two revealed its great extent. In 1840–41 the Australian Edward John Eyre travelled along the south coast from Adelaide to Albany, a distance of more than 1,300 miles (2,100 kilometres); the Australians Robert Burke and William John Wills travelled from Melbourne in the southeast to the Gulf of Carpentaria in the north.

Polar regions. The exploration of the polar regions was the work of the first half of the 20th century. Scientific curiosity mainly inspired the various enterprises, although political rivalry also played some part.

In the North Polar regions, the scientific age began with the voyaging of William Scoresby, an English whaler and scientist, who in 1806 reached 81°21′N. In 1828 an English explorer, Sir William Parry, travelling over drift ice from Svalbard, reached 82° N. The Norwegian explorer Fridtjof Nansen in 1893 attempted to reach the Pole by allowing his ship, the "Fram," to be frozen into the ice in the East Siberian Sea in the hope that a current would carry it over the Pole to east Greenland. At 84° N 102° E, Nansen with a companion left the ship and travelled by sled to 86°13′ N: the ship eventually emerged from the pack ice north of Svalbard. In 1909 an American explorer, Robert Peary, reached the North Pole by journeying by sled with 50 Eskimos from Ellesmere Island, northwest of Greenland. Soundings of 9,000 feet (2,700 metres) were made within five miles (eight kilometres) of the Pole; it seemed, therefore, that there could be no continent here. In 1958 the U.S. submarines "Skate" and "Nautilus" travelled across the Arctic Ocean under the ice cap.

Antarctica

The great southern continent, which Captain Cook demonstrated could not lie in the South Pacific, lay there neglected for some 50 years. From 1839 to 1843, the British rear admiral James Ross, in command of the ships "Erebus" and "Terror," explored the coast of Victoria Land. In 1894 Leonard Christensen, captain of a Norwegian whaler, landed a party at Cape Adare, the first to set foot on Antarctica. In the first decade of the 20th century, various explorers, including Britons such as William Bruce, Robert Falcon Scott, and Sir Ernest Henry Shackleton, the German Erich von Drygalski, and the Frenchman Jean-Baptiste Charcot, confirmed the existence of an ice cap of continental dimensions. In 1908–09 Shackleton led a brilliant expedition, during which he examined the Great Barrier, climbed to 11,000 feet (3,400 metres), and reached 88°23′ S. Scott and his party reached the Pole on January 17, 1912, only to find that the Norwegian explorer Roald Amundsen had already been there on December 14, 1911; Scott's party, caught in a blizzard, died on their return journey. In 1928 Sir Hubert Wilkins, the British explorer and aviator, flew over Grahamland, using Deception Island as a base. In 1957 and 1958 the British explorer Vivian Fuchs and Sir Edmund Hillary, the New Zealand mountaineer, travelled across the continent.

(J.B.Mi.)

EUROPEAN COLONIZATION

The age of modern colonialism began about 1500, following the European discoveries of a sea route around Africa's southern coast (1488) and of America (1492). With these events sea power shifted from the Mediterranean to the Atlantic and to the emerging nation-states of Portugal, Spain, the Dutch Republic, France, and England. By discovery, conquest, and settlement, these nations expanded and colonized throughout the world, spreading European institutions and culture.

European expansion before 1763

ANTECEDENTS OF EUROPEAN EXPANSION

Early communications with the Near East

Medieval Europe was largely self-contained until the First Crusade (1096–99), which opened new political and commercial communications with the Muslim Near East. Although Christian crusading states founded in Palestine and Syria proved ephemeral, commercial relations continued, and the European end of this trade fell largely into the hands of Italian cities.

Early European trade with Asia. The Oriental land and sea routes terminated at ports in the Crimea, until 1461 at Trebizond (now Trabzon, Turkey), Constantinople (now Istanbul), Asiatic Tripoli (in modern Lebanon), Antioch (in modern Turkey), Beirut (in modern Lebanon), and Alexandria (Egypt), where Italian galleys exchanged European for Eastern products.

Competition between Mediterranean nations for control of Asiatic commerce gradually narrowed to a contest between Venice and Genoa, with the former winning when it severely defeated its rival city in 1380; thereafter, in partnership with Egypt, Venice principally dominated the Oriental trade coming via the Indian Ocean and Red Sea to Alexandria.

Overland routes were not wholly closed, but the conquests of the central Asian warrior Timur (Tamerlane)—whose empire broke into warring fragments after his death in 1405—and the advantages of a nearly continuous sea voyage from the Middle and Far East to the Mediterranean gave Venice a virtual monopoly of some Oriental products, principally spices. The word spices then had a loose application and extended to many Oriental luxuries, but the most valuable European imports were pepper, nutmeg, cloves, and cinnamon.

The Venetians distributed these expensive condiments throughout the Mediterranean region and northern Europe; they were shipped to the latter first by pack trains up the Rhône Valley and, after 1314, by Flanders' galleys to the Low Countries, western Germany, France, and England. The fall of Constantinople to the Ottoman Turks in 1453 did not seriously affect Venetian control. Although other Europeans resented this dominance of the trade, even the Portuguese discovery and exploitation of the Cape of Good Hope route could not altogether break it.

Early Renaissance Europe was short of cash money, though it had substantial banks in northern Italy and southern Germany. Florence possessed aggregations of capital, and its Bardi bank in the 14th century and the Medici successor in the 15th financed much of the eastern Mediterranean trade.

Later, during the great discoveries, the Augsburg houses of Fugger and Welser furnished capital for voyages and New World enterprises.

Gold came from Central Africa by Saharan caravan from Upper Volta (Burkina Faso) near the Niger, and interested persons in Portugal knew something of this. When Prince Henry the Navigator undertook sponsorship of Portuguese discovery voyages down the west coast of Africa, a principal motive was to find the mouth of a river to be ascended to these mines.

Technological improvements. Europe had made some progress in discovery before the main age of exploration. The discoveries of the Madeira Islands and the Azores in the 14th century by Genoese seamen could not be followed up immediately, however, because they had been made in galleys built for the Mediterranean and ill suited to ocean travel; the numerous rowers that they required and their lack of substantial holds left only limited room for provisions and cargo. In the early 15th century all-sails vessels, the caravels, largely superseded galleys for Atlantic travel; these were light ships, having usually two but sometimes three masts, ordinarily equipped with lateen sails but occasionally square-rigged. When longer voyages began, the *nao,* or carrack, proved better than the caravel; it had three masts and square rigging and was a rounder, heavier ship, more fitted to cope with ocean winds.

Navigational instruments were improved. The compass, probably imported in primitive form from the Orient, was gradually developed until, by the 15th century, European pilots were using an iron pin that pivoted in a round box. They realized that it did not point to the true north, and no one at that time knew of the magnetic pole, but they learned approximately how to correct the readings. The astrolabe, used for determining latitude by the altitude of stars, had been known since Roman times, but its employment by seafarers was rare, even as late as 1300; it became more common during the next 50 years, though most pilots probably did not possess it and often did not need it because most voyages took place in the narrow waters of the Mediterranean or Baltic or along western European coasts. For longitude, then and many years thereafter, dead reckoning had to be employed, but this could be reasonably accurate when done by experts.

The typical medieval map had been the planisphere, or

Medieval maps

mappemonde, which arranged the three known continents in circular form on a disk surface and illustrated a concept more theological than geographical. The earliest surviving specimens of the portolanic, or harbour-finding, charts date from shortly before 1300 and are of Pisan and Genoese origin. Portolanic maps aided voyagers by showing Mediterranean coastlines with remarkable accuracy, but they gave no attention to hinterlands. As Atlantic sailings increased, the coasts of western Europe and Africa south of the Strait of Gibraltar were shown somewhat correctly, though less so than for the Mediterranean.

THE FIRST EUROPEAN EMPIRES (16TH CENTURY)

Portuguese dominance of Eastern commerce

Portugal's seaborne empire. Following Christopher Columbus' first voyage, the rulers of Portugal and Spain, by the Treaty of Tordesillas (1494), partitioned the non-Christian world between them by an imaginary line in the Atlantic, 370 leagues (about 1,300 miles) west of the Cape Verde Islands. Portugal could claim and occupy everything to the east of the line and Spain everything to the west (though no one then knew where the demarcation would bisect the other side of the globe). Portuguese rule in India, the East Indies, and Brazil rested on this treaty, as well as on Portuguese discoveries and on papal sanction (Pope Leo X, by a bull of 1514, forbade others to interfere with Portugal's possessions). Except for such minor incursions as those of Ferdinand Magellan's surviving ship in 1522 and the Englishman Sir Francis Drake's voyage around the world in 1577–80, the Portuguese operated in the East for nearly a century without European competition. They faced occasional Oriental enemies but weathered these dangers with their superior ships, gunnery, and seamanship.

Territorially, theirs was scarcely an empire; it was a commercial operation based on possession of fortifications and posts strategically situated for trade. This policy was carried out principally by two viceroys, Francisco de Almeida in 1505–09 and Afonso de Albuquerque in 1509–15. Almeida seized several eastern African and Indian points and defeated a Muslim naval coalition off Diu (now in Goa, Daman, and Diu union territory, India). Albuquerque endeavoured to gain a monopoly of European spice trade for his country by sealing off all entrances and exits of the Indian Ocean competing with the Portuguese route around the Cape of Good Hope. In 1510 he took Goa, in western India, which became the capital and stronghold of the Portuguese East, and in 1511 he captured Malacca at the farther end of the ocean. Later he subdued Hormuz (now in Iran), commanding the Persian Gulf. They brought soldiers from the home country in limited numbers; but the Portuguese also relied on alliances with native states and enlisted sepoy troops, a policy later followed by the French and English.

Portugal never fully dominated the Indian Ocean because it lacked warships necessary to control the vast water expanse. Albuquerque's failure to capture Aden at the Red Sea entrance allowed the old traffic through Egypt to Venice to resume following an initial dislocation, and this continued after the Ottoman Turks conquered Egypt in 1517. Much of the Indian Ocean trade was local and, until the Portuguese incursion, had been conducted by Arabs or at least by Muslims. The Portuguese, who at first had intended to oust the Arabs entirely, found it impossible to manage without them. The Hindus, whom they hoped to use for local trade purposes, proved unenterprising and had caste restrictions regarding sea voyages. Muslims were soon trafficking again vigorously, with Portuguese sanction.

Portuguese subjects also pressed beyond the Strait of Malacca to the East Indies, Siam (now Thailand), and Canton in Ming-dynasty China. Trade with the celestial empire, difficult at first because of China's exclusionist policies, at length grew, especially after Portugal in 1557 leased Macau, through which for the next 300 years passed much of the Occidental trade with China. Individual Portuguese reached Japan in 1542, followed by traders and Francis Xavier (later made a saint), a renowned Jesuit missionary who laboured with small success to make converts. In the 17th century, the Japanese adopted a rigorous exclusionist policy, although they allowed Portugal's successors, the Dutch, to conduct a limited trade from the small island of Deshima, near Nagasaki.

Partial domination of the Indian Ocean and much of its valuable trade did not bring Portugal's crown as much profit as had been anticipated. The intention had been to make Oriental trade a royal monopoly; but Portuguese, from viceroys to humble soldiers and seamen, became private merchants and lined their own pockets to the deprivation of the royal treasury. The Eastern footholds were expensive to maintain, and frequent mishaps to vessels of the Indian fleets, from shipwreck or enemies, reduced gains. The lack of a true monopoly prevented the Portuguese from charging the prices that they wished in European markets. Moreover, Lisbon, while an ideal starting point for voyages around the Cape, proved poorly situated as a distribution centre for spice to northern and central Europe. Antwerp, on the Scheldt, was far superior, and for a time Portugal maintained a trading house there; but Portuguese agents found spice sales taken out of their hands by more experienced Italian, German, and Flemish merchants, and the Antwerp establishment was closed in 1549.

It has been asserted that the Portuguese had no racial prejudice, but their record proves the opposite. In the 16th and 17th centuries, they could not be expected to be tolerant of Oriental religions, although they soon recognized that wholesale conversion to Catholicism was impossible. Some Africans and Asiatics became Christians and even entered the clergy; but seldom if ever did they rise above the status of parish priests. In other affairs the Portuguese generally treated the dark-skinned peoples as inferiors.

The east coast of Brazil belonged to Portugal by the Tordesillas pact. The government of Manuel I and his successor, John III (ruled 1521–57), paid it small attention for 30 years. It proved nearly useless as a way station to the Cape; its Indian population was savage, and its products, consisting chiefly of *pau-brasil* (Brazilian dyewood), yielded much less revenue than those of India. Threats of French and Spanish intrusion caused John III, in 1530, to send Martim Afonso de Sousa to make a careful survey of the Brazilian coast and to suggest sites for colonization. Next, the littoral was partitioned into strips called *capitanias,* each colonized and governed under feudal terms by a proprietor, or *donatário.* Some limited settlement followed, and in 1549 the *capitanias* were united under a governor general who established residence at Bahia (now Salvador, Brazil).

Portuguese colonization of Brazil

In 1580 Philip II of Spain seized the Portuguese throne, which had fallen vacant and to which he had some blood claim. Portugal remained theoretically independent, bound only by a personal union to its neighbour; but succeeding Spanish monarchs steadily encroached on its liberties until the small kingdom became, in effect, a conquered province. Spain's European enemies meanwhile descended on the Portuguese Empire and ended its Eastern supremacy before the restoration of Portugal's independence in 1640.

Beginnings of the Spanish conquest in America

Spain's American empire. *The conquests.* Only gradually did the Spaniards realize the possibilities of America. They had completed the occupation of the larger West Indian islands by 1512, though they largely ignored the smaller ones, to their ultimate regret. Thus far they had found lands nearly empty of treasure, populated by naked primitives who died off rapidly on contact with Europeans. In 1508 an expedition did leave Hispaniola to colonize the mainland, and, after hardship and decimation, the remnant settled at Darién on the Isthmus of Panama, from which in 1513 Vasco Núñez de Balboa made his famous march to the Pacific. On the Isthmus the Spaniards heard garbled reports of the wealth and splendour of Inca Peru. Balboa was succeeded (and judicially murdered) by Pedrarias Dávila, who turned his attention to Central America and founded Nicaragua.

Expeditions sent by Diego Velázquez, governor of Cuba, made contact with the decayed Mayan civilization of Yucatán and brought news of the cities and precious metals of Aztec Mexico. Hernán Cortés entered Mexico from Cuba in 1519 and spent two years overthrowing the Aztec confederation, which dominated Mexico's civilized heart-

land. The Spaniards used firearms effectively but did most of their fighting with pikes and blades, aided by numerous Indian allies who hated the dominant Aztecs. The conquest of Aztec Mexico led directly to that of Guatemala and about half of Yucatán, whose geography and warlike inhabitants slowed Spanish progress.

Mexico yielded much gold and silver, and the conquerors imagined still greater wealth and wonders to the north. None of this existed, but it seemed real when a northern wanderer, Alvar Núñez Cabeza de Vaca, in 1536 brought to Mexico an exciting but fanciful report of the fabulous lands. Expeditions explored northern Mexico and the southern part of what is now the United States—notably the expedition of Juan Rodríguez Cabrillo by sea along what are now the California and Oregon coasts and the expeditions of Hernando de Soto and Francisco Vázquez Coronado through the southeastern and southwestern U.S. regions. These brought geographical knowledge but nothing of value to the Spaniards, who for years thereafter ignored the northern regions.

Meanwhile, the Pizarro brothers—Francisco Pizarro and his half-brothers Gonzalo and Hernando—entered the Inca Empire from Panama in 1531 and proceeded with its conquest. Finding the huge realm divided by a recent civil war over the throne, they captured and executed the incumbent usurper, Atahualpa. But the conquest took years to complete; the Pizarros had to crush a formidable native rising and to defeat their erstwhile associate, Diego de Almagro, who felt cheated of his fair share of the spoils. The Pizarros and their followers took and divided a great amount of gold and silver, with prospects of more from the mines of Peru and Bolivia. By-products of the Inca conquest were the seizure of northern Chile by Pedro de Valdivia and the descent of the entire Amazon by Francisco de Orellana. Other conquistadors entered the regions of what became Ecuador, Colombia, and Argentina. (See LATIN AMERICA, THE HISTORY OF.)

A colonial period of nearly three centuries followed the major Spanish conquests. The empire was created in a time of rising European absolutism, which flourished in both Spain and Spanish America and reached its height in the 18th century. The overseas colonies became and remained the king's private estate.

Spanish colonial policies. Shortly before the death of Queen Isabella I in 1504, the Spanish sovereigns created the House of Trade (Casa de Contratación) to regulate commerce between Spain and the New World. Their purpose was to make the trade monopolistic and thus pour the maximum amount of bullion into the royal treasury. This policy, seemingly successful at first, fell short later because Spain failed to provide necessary manufactured goods for its colonies, foreign competitors appeared, and smuggling grew.

In 1524 Charles V created the Council of the Indies (Consejo de Indias) as a lawmaking body for the colonies. During the three centuries of its existence, this council enacted a massive amount of legislation, though much grew obsolete and became a dead letter. The industrious Philip II died in 1598, and his indolent or incompetent successors left American affairs to the Casa and Consejo; both proved generally conscientious and hard-working bodies, though, for a time in the 17th century, appointments to the legislating council could be purchased.

The viceregal system

The viceregal system dated from 1535, when Antonio de Mendoza was sent to govern New Spain, or Mexico, bypassing the still-vigorous Cortés. A second viceroy was named for Peru in 1542, and the viceroyalties of New Granada and Río de la Plata were formed in 1739 and 1776, respectively. By the 18th century, viceroys served average terms of five years, and under them functioned a hierarchy of bureaucrats, nearly all sent from Spain to occupy frequently lucrative posts. American-born Spaniards resented this favouritism shown the peninsular Spaniards, and their jealousy accounted in part for their later separation from Spain. Lower socially and economically than either white class were the mestizo offspring of white and Indian matings, and still lower were the Indians and black slaves.

Though a belief to the contrary exists, Spain sent many colonists to America. One indication of this is the number of new cities founded, distinct from the old Indian culture centres. A partial list of such cities, besides the early island ones, includes Vera Cruz, New Spain; Panama, Cartagena, and Guayaquil, in New Granada (in modern Panama, Colombia, and Ecuador, respectively); Lima, Peru; and all those of what are now Chile, Paraguay, Argentina, and Uruguay.

A problem early faced and never truly solved by Spain was that of the Indians. The home government was generally benevolent in legislating for their welfare but could not altogether enforce its humane policies in distant America. The foremost controversy in early decades involved the *encomienda,* by which Indian groups were entrusted to Spanish proprietors, who in theory cared for them physically and spiritually in return for rights to tribute and labour but who in practice often abused and enslaved them.

Spanish Dominican friars were the first to condemn the *encomienda* and work for its abolition; the outstanding reformer was a missionary, Bartolomé de Las Casas, who devoted most of his long life to the Indian cause. He secured passage of laws in 1542 ordering the early abolition of the *encomienda,* but efforts to enforce these brought noncompliance in New Spain and armed rebellion in Peru. A belief held by some Spanish theologians—that Indians were inferior beings who were destined to be natural slaves, to be subdued and forcibly converted to Christianity—generally prevailed over the opposition of Las Casas and fellow Dominicans. The *encomienda* or its equivalent endured, although this feudal institution declined as royal absolutism grew.

Diminution of the Indians

The Indians became real or nominal Christians, but their numbers shrank, less from slaughter and exploitation than from Old World diseases, frequently smallpox, for which they had no inherited immunity. The aboriginal West Indian population virtually disappeared in a few generations, to be replaced by black slaves. Indian numbers shrank in all mainland areas: at the beginning of Spanish settlement there were perhaps 50,000,000 aborigines; the figure had decreased to an estimated 4,000,000 in the 17th century, after which it slowly rose again. Meanwhile the hybrid mestizo element grew and—to a limited extent—replaced the Indians.

The Leyenda Negra (Black Legend) propagated by critics of Spanish policy still contributes to the general belief that Spain exceeded other nations in cruelty to subject populations; on the other hand, a review of Spain's record suggests that it was no worse than other nations and, in fact, produced a greater number of humanitarian reformers. When Dominican zeal declined, the new and powerful Jesuit order became the major Indian protector and led in missionary activity until its expulsion from the Spanish Empire in 1767; the Jesuits took charge of large converted native communities, notably in the area of the viceroyalty of Río de la Plata that is now Paraguay, in their paternalism often imposing stern discipline.

Effects of the discoveries and empires. Before the discovery of America and the sea route to Asia, the Mediterranean had been the trading and naval centre of Europe and the Near East. Italian seamen were rightly considered to be the best, and they commanded the first royally sponsored transatlantic expeditions—Columbus for Spain, John Cabot for England, and Giovanni da Verrazano for France.

Europe's shift to the Atlantic. Until then the Western countries had lain on the fringe of civilization, with nothing apparently beyond them but Iceland and small islands. With the discovery of the Cape route and America, nations formerly peripheral found themselves central, with geographical forces impelling them to leadership.

The Mediterranean did not become a backwater, and the Venetian republic remained a major commercial power in the 16th century. Venice's decline came in the 17th, though the Venetians were still formidable against the Turks. As the more powerful Dutch, French, and English replaced the Eastern pioneers of Portugal, however, the burden of competition became more than the venerable republic could bear. The last decisive naval battle fought

wholly by Mediterranean seamen was Lepanto (Náupaktos, Greece), where Don John of Austria, in 1571, commanding Spanish and Italian galleys, defeated an Ottoman fleet. Although Atlantic powers thereafter often fought in the Mediterranean, they mainly fought each other, while the Italian cities became pawns in international politics. The nation-state was superseding the small principality and city-state, a trend that had begun before the discoveries. The new nations lay on the Atlantic; and, though Spain and France had Mediterranean frontages, the advantage went to those seaports belonging to substantial countries with ready access to the outer world.

Changes in Europe. The opening of old lands in Asia and new ones in America changed Europe forever, and the Iberian countries understandably felt the changes first. The Portuguese government, for a time, made large profits from its Eastern trade, and individuals prospered; but Oriental luxuries were costly compared with the European goods that Portugal offered, and the balance had to be made up in specie. This eastward drain of gold and silver had gone on long before Portuguese imperial times, but it was now intensified. Much of the bullion reaching the Orient did not circulate but was hoarded or made into ornaments; consequently, there was no inflation in Asia, and prices there did not rise enough to create a demand for Western goods, which would have reversed the flow of bullion from the West. The Portuguese obtained most of the precious metal for this trade from spice sales through Antwerp and from Africa. The drain proved critical, and, by the reign of John III, the government found itself hard pressed economically and forced to abandon overseas posts that were a financial burden. Later, beginning in the 17th century, Portugal drew its own supply of jewels and gold from Brazil.

Eastward drain of Portuguese gold and silver

Spain's case was the reverse; although the first American lands discovered yielded little mineral wealth, the mines of Mexico by the 1520s and those of Potosí (in modern Bolivia) by the 1540s were shipping to Spain large quantities of bullion, much of it crown revenue. This did not furnish Charles V and Philip II their largest income; Spanish taxation still exceeded wealth from the New World, yet American silver and gold proved sufficient to cause a price revolution in Spain, where costs, depending on the region, were multiplied by three and five during the 16th century. The Spanish government wished to keep bullion from leaving the kingdom, but high prices in Spain made it a good market for outside products. Spanish industry declined in the 16th century, in part because of the sales taxes imposed by the crown, which necessitated more buying of foreign merchandise. Great quantities of bullion had to be poured out to finance the expensive Spanish European empire and the costly wars and diplomacy of Charles V and Philip II, both of whom were constantly in debt.

Price rises followed in other countries, largely from the influx of Spanish bullion. In England, where some statistics are available, costs by 1650 had risen by 250 percent over those of 1500.

The European commercial revolution, which brought increased industry, more trade, and larger banks, had begun before the discoveries, but it received stimulus from them. Bullion from America helped create a money economy, replacing the older and largely barter exchange—a trend accentuated by greater European mineral production in the early 16th century. The trade emporiums of Italy and the Baltic Hanseatic League declined and were largely replaced by those of the Dutch Republic, England, and France. Joint-stock companies made an impressive appearance, notably the East India Companies of the Dutch Republic, England, and France in the 17th century. The mercantile theory that precious metals constitute the true wealth, though it had attracted advocates for a long time, now came into full vogue and continued to dominate economic thinking.

New products for Europe

Discovery introduced Europe to new foods and beverages. Coffee, from Ethiopia, had been consumed in Arabia and Egypt before its wide European use began in the 17th century. Tobacco, an American plant smoked by Indians, won an Old World market despite many individual objectors; the same proved true of chocolate from Mexico and tea from Asia. The South American potato became a staple food in such places as Ireland and central Europe. Cotton, from the Old World, took firm root in the New, from which Europe received an enormously increased supply. Sugar, introduced to the American tropics, along with its molasses and rum derivatives, in time became the principal exports of those regions. Spice was certainly more plentiful than before the discoveries, though the Dutch, when they controlled the East Indies, were able to limit production and thus to keep the price of cloves and nutmegs high.

The influence of the discoveries permeated literature. Sir Thomas More's *Utopia,* printed in 1516 and dealing with an imaginary island, was suggested by South America. The Portuguese poet Luís de Camões recounted the voyage of Vasco da Gama, though fancifully, in epic verse. Michel de Montaigne discoursed upon American savages, some of whom he had seen in France. Christopher Marlowe's drama *Tamburlaine* (1587), though based on the life of the Asiatic conqueror, was an exhortation to his fellow Englishmen to penetrate the New World.

Historiography acquired a broader base by taking the newly discovered lands into account. Astronomy was revolutionized by European penetration of the Southern Hemisphere and discovery of constellations unknown before. Map makers, typified by the Fleming Gerardus Mercator and the Dutchman Abraham Ortelius, portrayed the world in terms that are still recognizable.

COLONIES FROM NORTHERN EUROPE AND MERCANTILISM (17TH CENTURY)

The northern Atlantic powers, for understandable reasons, acquired no permanent overseas possessions before 1600. The United Provinces of the Netherlands spent the final decades of the 16th century winning independence from Spain; France had constant European involvements and wars of religion; England, matrimonially allied with Spain as late as 1558, was undergoing its Protestant Reformation and long was unwilling to challenge predominant Spain openly in any manner.

The Dutch. Although England's defeat of Philip II's Armada in 1588 helped to lessen Spanish sea power, it was the Dutch who early in the next century really broke that power and became the world's foremost naval and commercial nation, with science and skills commensurate with their prowess. Only late in the 17th century did they decline, because of Holland's limited size and the inferiority of its geographical position to England's. The Dutch, meanwhile, penetrated all the known oceans, including the Arctic, and waged unrelenting war against the Iberian kingdoms.

The Dutch coveted the Portuguese commercial empire more than the Spanish continental one. They took much of the Portuguese East and invaded Brazil (1624–54), the richer half of which they controlled for a time. They also penetrated Portuguese Angola, which they desired because the slaves it exported were beginning to work the Brazilian plantations. They ultimately failed in the South Atlantic, though they gained Dutch Guiana (now Suriname), Curaçao, and what later became British Guiana (Guyana). Meanwhile, Willem Schouten, one of their free-lance voyagers, had made the discovery of Cape Horn in 1616.

Eastern pursuits. The Dutch States-General, in 1602, chartered the United East India Company (Vereenigde Oost-Indische Compagnie, popularly called the Dutch East India Company), a joint-stock enterprise with investment open to all. In control was a board of 17 directors, the so-called Heeren XVII, who received a monopoly of navigational rights eastward around the Cape of Good Hope and westward through the Strait of Magellan. They could make treaties with native princes on behalf of the States-General (from which they were scarcely separable), establish garrisoned forts, and appoint governors and justices. The company had no interest in extending Protestantism, and there was no mention of religious conversion, though Calvinist ministers later gained converts in the East, mostly in communities previously made Catholic by Portuguese Jesuits.

Formation of the United East India Company

The company established headquarters first at Bantam in Java in 1607, later moving them to Jacatra, renamed Batavia (now Jakarta), in the same island. Its two main objectives were the ouster of European competitors—Portuguese, English, and Spanish—and dominance of local trade, previously in native hands. Portuguese vigour had somewhat declined, and the Dutch were victorious in most armed encounters. They also squeezed out the English, whose own East India Company thereafter concentrated efforts in the Indian peninsula.

The principal builder of the Dutch Oriental empire was Jan Pieterszoon Coen, company governor general from 1618 to 1623 and again from 1627 until his death in 1629. Financially, local trade monopoly was even more important than the expulsion of white competitors. The extension of Dutch control to islands beyond Java had started before the governorship of Coen, who accentuated the process. He and other company officials behaved ruthlessly; for example, when the inhabitants of the nutmeg-growing island of Great Banda (modern Pulau Banda Besar in Indonesia) resisted the Dutch in 1621, Coen had 2,500 of the inhabitants massacred and 800 more transported to Batavia. Company policy was to restrict clove production to Amboina and a few neighbouring islands firmly under Dutch control. To insure this, about 65,000 clove trees were destroyed in the Moluccas, and Dutch subjection of Macassar made the monopoly virtually complete. In 1656 the famous Moluccas were described as a wilderness. Besides being a conqueror, Coen was an able businessman and an economist. When he died he was engaged in gaining a monopoly of the pepper of interior Sumatra, which was later sealed off securely by the fall of Portuguese Malacca in 1641.

Batavia as the focal point of the Dutch East

Batavia became the focal point of the Dutch East, and through it passed the commerce of China, Japan, India, Ceylon, and Persia, bound for Europe or other Oriental ports. The Dutch never monopolized the China trade because the Portuguese held Macau, the Spaniards held Manila, and the Japanese, for a time, engaged in this commerce. The Dutch gained a foothold in Formosa in 1624 but lost it to a Chinese pirate in 1662. After Japan became exclusionist in 1641, a trickle of Dutch trade continued to enter it through the small island of Deshima (now part of Nagasaki, Japan), even after the dissolution of the United East India Company in 1799.

The economy of Java changed somewhat after the importation of the coffee plant in 1696. Coffee, often simply called java, rapidly became a major island crop and was exported from there to Dutch America. The company had earlier brought coffee to Ceylon (now Sri Lanka), but that experiment had failed when a blight attacked its leaves. The company ousted the Portuguese from Ceylon and dominated the island until it was itself dispossessed by the British in 1796. Under its jurisdiction, as earlier, the major Ceylonese export was cinnamon, though the Dutch also dealt in jewels and pepper and carried on a trade in elephants.

In their constant search for commercial outlets, the company's officials sponsored new exploration. Coen's ablest successor, Antonio van Diemen, governor general in 1636–45, sent Abel Tasman to investigate the great land (Australia) previously sighted by Spanish, Portuguese, and Dutch seamen. Tasman sailed around the continent and discovered Van Diemen's Land (Tasmania), Staatenland (New Zealand), and the Tonga and Fiji Islands, but their commercial possibilities seemed insufficient to warrant further attention.

Dutch penetration of the East was not colonization; small farmers and artisans neither could nor would compete with the abundant, cheap native labour. Those Dutchmen going eastward were company officials, seamen and soldiers, overseers of plantations and commerce, and a few scientists and Calvinist clergymen; there was no place for others.

The Dutch moved into uninhabited Mauritius, which they later abandoned and saw pass first to France and finally to Great Britain. The Heeren XVII felt the need of a station on the arduous voyage between the home country and the East. They obtained it at Cape Town (founded in 1652 by Jan van Riebeeck), which company ships thereafter regularly visited for fresh meat and vegetables to reduce scurvy. The town did not altogether live up to first expectations because the harbour was exposed, but the hinterland possessed a good climate and no dangerous natives. Beginning in the 1680s the company encouraged a moderate influx by Dutch families and French Huguenot exiles. Although the British conquered the colony in 1806, the descendants of these early settlers remained the largest white element and spoke a variant of Dutch, which became Afrikaans.

Western pursuits. Dutch activity in the South Atlantic, Guyana, the West Indies, and New Netherland (New York) was the work of the West India Company (West-Indische Compagnie), founded in 1621. This never proved as successful as the Heeren XVII's generally profitable enterprise, but it did produce results. Except for the Cape, the only real Dutch colonization undertaking was New Netherland in North America, started in 1624 by the West India Company. Ft. Amsterdam, or New Amsterdam, was founded, and two years later the company agent Peter Minuit made a 60-guilder ($24) transaction with the local Indians for the purchase of Manhattan island. Dutch settlement along the Hudson from New Amsterdam to Ft. Orange (Albany) remained sparse; the company's insistence on monopolizing the Indian fur trade discouraged Dutchmen from migrating there. Further, the policy of creating large patroon land grants, five in all, along the river under feudal proprietors, limited settlement. New Amsterdam itself became fairly thriving because it possessed the best harbour in North America. Many besides Dutchmen settled there; some came from nearby New England, and there was a sprinkling of French, Scandinavian, Irish, German, and Jewish inhabitants. The city was weakly defended and fell rather easily to an English fleet in 1664; it was renamed New York. Although the Dutch retook it briefly in 1673–74, the colony became permanently English by the Treaty of Westminster in 1674. The West India Company was then dissolved, to be reconstituted for exploitation of the Caribbean holdings but to attempt no further territorial expansion.

Founding of the Dutch West India Company

The French. France probably could have become the leading European colonial power in the 17th and 18th centuries. It had the largest population and wealth, the best army while Louis XIV ruled, and, for a time in his reign, the strongest navy. But France pursued a spasmodic overseas policy because of an intense preoccupation with European affairs; England, France's ultimately successful rival, was freer of such entanglements.

Early settlements in the New World. Verrazano reconnoitered the North American coast for France in 1524, and in the next decade Jacques Cartier explored the St. Lawrence River; his plans to establish a colony, however, came to nothing. During most of the rest of the 16th century, French colonization efforts were confined to short-lived settlements at Guanabara Bay (Rio de Janeiro) and Florida; both met sad ends. France meanwhile was troubled by internal religious strife and, for a time, was influenced by Philip II of Spain. But at the beginning of the 17th century, with Spanish power declining and domestic religious peace restored by King Henry IV's Edict of Nantes (1598), granting religious liberty to the Huguenots, the King chartered a Compagnie d'Occident (Western Company). This led to further exploration and to a small Acadian (Nova Scotian) settlement, and in 1603 Samuel de Champlain went to Canada, called New France. Champlain became Canada's outstanding leader, founding Quebec in 1608, defeating the Iroquois of New York, stimulating fur trade, and exploring westward to Lake Huron in 1615. He introduced Recollet (Franciscan) friars for conversion of the American Indians, but the Jesuit order (the Society of Jesus) soon became the principal missionary body in Canada.

Under the ministership of Cardinal Richelieu (served 1624–42), a Council of Marine was created, with responsibility for colonial affairs. French West Indian settlement, following the activities of pirates and filibusters, began in 1625 with the admission of French settlers to St. Christopher (already settled by the British in 1623 and partitioned

French West Indian settlement

between the two countries until its cession to the British in 1713), and by 1664 France held 14 Antillean islands containing 7,000 whites, the principal possessions being Guadeloupe and Martinique. Saint-Domingue (Haiti), not yet annexed, contained numbers of Frenchmen, mostly buccaneers from Tortuga. Sugar became the main crop of the islands; the date when importation of black slaves began is uncertain, though some were sold at Guadeloupe as early as 1642.

French West Indian society was caste bound, with officials and large planters (*gros blancs*) at the top, followed, in descending order, by merchants, buccaneers, and small farmers (*petits blancs*). Lowest of all were contract labourers from France (*engagés*) and black slaves.

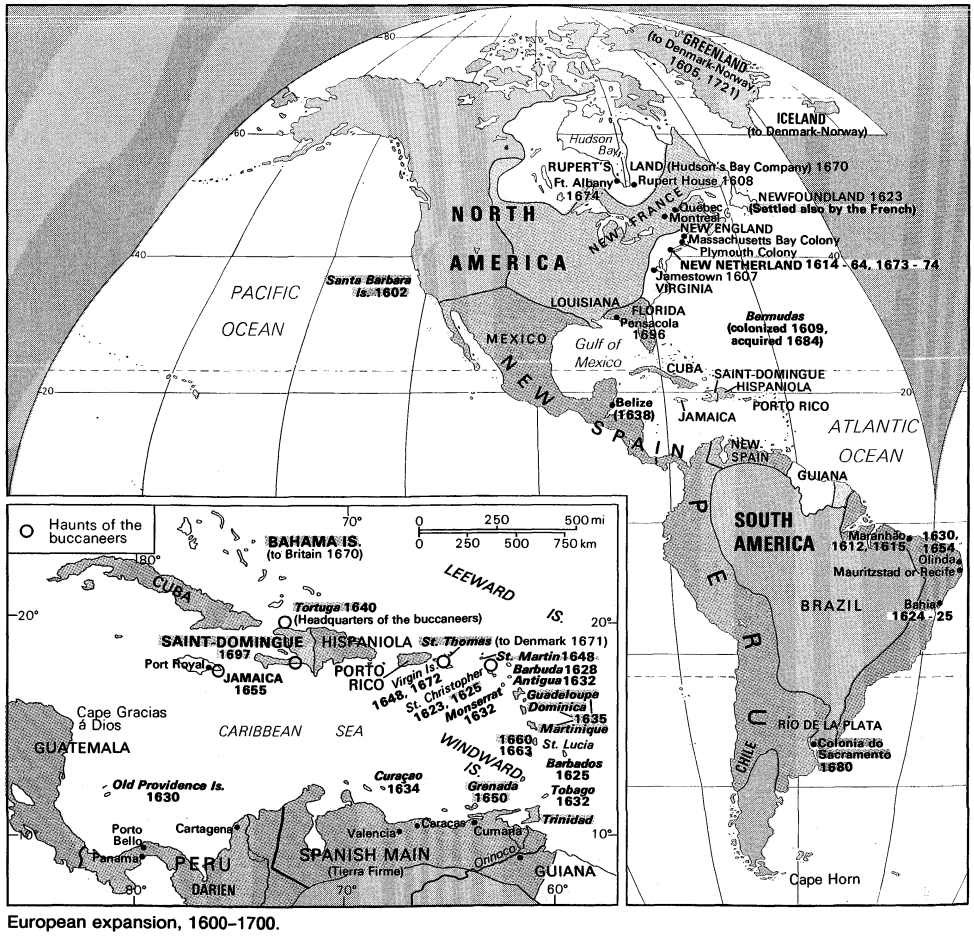
French Guiana was built around the Cayenne settlement, founded about 1637. There were other Frenchmen along the neighbouring coast at first, but, threatened by Dutchmen and natives, they finally took refuge at Cayenne. The Cayenne settlers, lacking any basis of prosperity, existed partly by raiding the Amazon Indians. The 18th century brought some improvement, but as late as 1743 French Guiana had only 600 whites, living by coffee and cacao culture and without means to import any but the crudest necessities.

Activities in India. Jean-Baptiste Colbert held a succession of high offices in France, including the ministry of marine, during the early reign of Louis XIV. Colbert was an archmercantilist and believed that an abundance of precious metals would enrich France. This required a favourable balance of trade and protective tariffs. Most of his policy applied to France itself, but he meant to supplement it with colonial markets protected by a strong navy. Colbert felt concern over the quantities of cash that Frenchmen paid the Dutch for Eastern products and intended for his countrymen to have a share of those profits. In 1664 he placed hopes in a new French Company of the East Indies (Compagnie Française des Indes Orientales), to which he personally subscribed and which bought out small predecessors. The company tried unsuccessfully to make Madagascar a great centre of trade, and the huge island became a stronghold of piracy, though the French acquired nearby Mauritius.

In the Indian peninsula, where the English East India Company had holdings, French progress was slow in Colbert's time and after, partly because the last great Mughal emperor, Aurangzeb, reigned and dominated India. The company did acquire Pondichéry and several other posts, however, and an affiliate opened a limited trade with China. When Aurangzeb died in 1707, his empire declined rapidly. Thereafter, the question of future control of India lay chiefly between the French company (reorganized and renamed the Compagnie Française des Indes in 1720) and the English company; both companies backed or opposed warring native rulers and exacted payment from them for financial support and for arming and drilling the native sepoy troops in the European manner. By the 1740s the French had gained the upper hand, and in the War of the Austrian Succession (1740–48; called King George's War in North America), the French governor general of India, Joseph-François Dupleix, captured Madras, the centre of British power. But in the ensuing Treaty of Aix-la-Chapelle the British, who had made gains in North America, recovered Madras. Never again did the French come so near success, and their fortunes soon declined. Their company had not made large profits because expensive wars and the costs of subsidizing native princes had consumed revenue. The home government seldom cooperated, and French investors on the whole declined to speculate in overseas ventures.

French and English rivalry

Colonization of New France. New France became a royal province in 1663, with both good and bad results. The arrival of troops in 1665 lessened the danger from the hostile Iroquois. Jean Talon, the powerful intendant sent by Colbert in the same year, strove to make Canada a self-sustaining economic structure, but his plan was finally thwarted by his home government's failure to supply fi-



European expansion, 1600–1700.

nancial means chiefly because of the King's extravagance and costly European wars.

Colbert gave some stimulus to colonization of New France. Grants of land, called *seigneuries,* with frontages on the St. Lawrence, were apportioned to proprietors, who then allotted holdings to small farmers, or habitants. More land came under cultivation, and the white population grew, though immigration from France declined sharply after 1681 because the home authorities were reluctant to spare manpower for empty Canada. After 1700 most French Canadians were North American born, a factor that weakened loyalty to the mother country.

North American exploration proceeded rapidly in Colbert's time. Fur traders had earlier reached Lake Superior; Louis Jolliet and Jacques Marquette now travelled the Fox and Wisconsin rivers to the Mississippi in 1673 and descended it to the Arkansas. Robert Cavelier, sieur de La Salle, followed the Mississippi to the Gulf of Mexico in 1682 and claimed the entire Mississippi River Basin, or Louisiana, for France; a later consequence was the founding of New Orleans (Nouvelle-Orléans) in 1718 by Jean-Baptiste Lemoyne, sieur de Bienville, the governor of Louisiana. French traders ultimately reached Santa Fe in Spanish New Mexico, and the sons of explorer Pierre Gaultier de Varennes, sieur de la Vérendrye—Louis-Joseph and François—visited the Black Hills of South Dakota and may have seen the Rocky Mountains.

The Roman Catholic Church became firmly rooted in Canada, without the intellectual opposition and anticlericalism that developed in 18th-century France. Jesuit mission work among the Indians, extending to the Middle West, saw more devotion and bravery by the priests than substantial results. Christianity made small appeal to most Indians, who could accept a supreme being but rejected the Christian ethic. Several zealous Jesuits became martyrs to the faith; genuine conversions were few and backslidings frequent.

Jesuit mission work

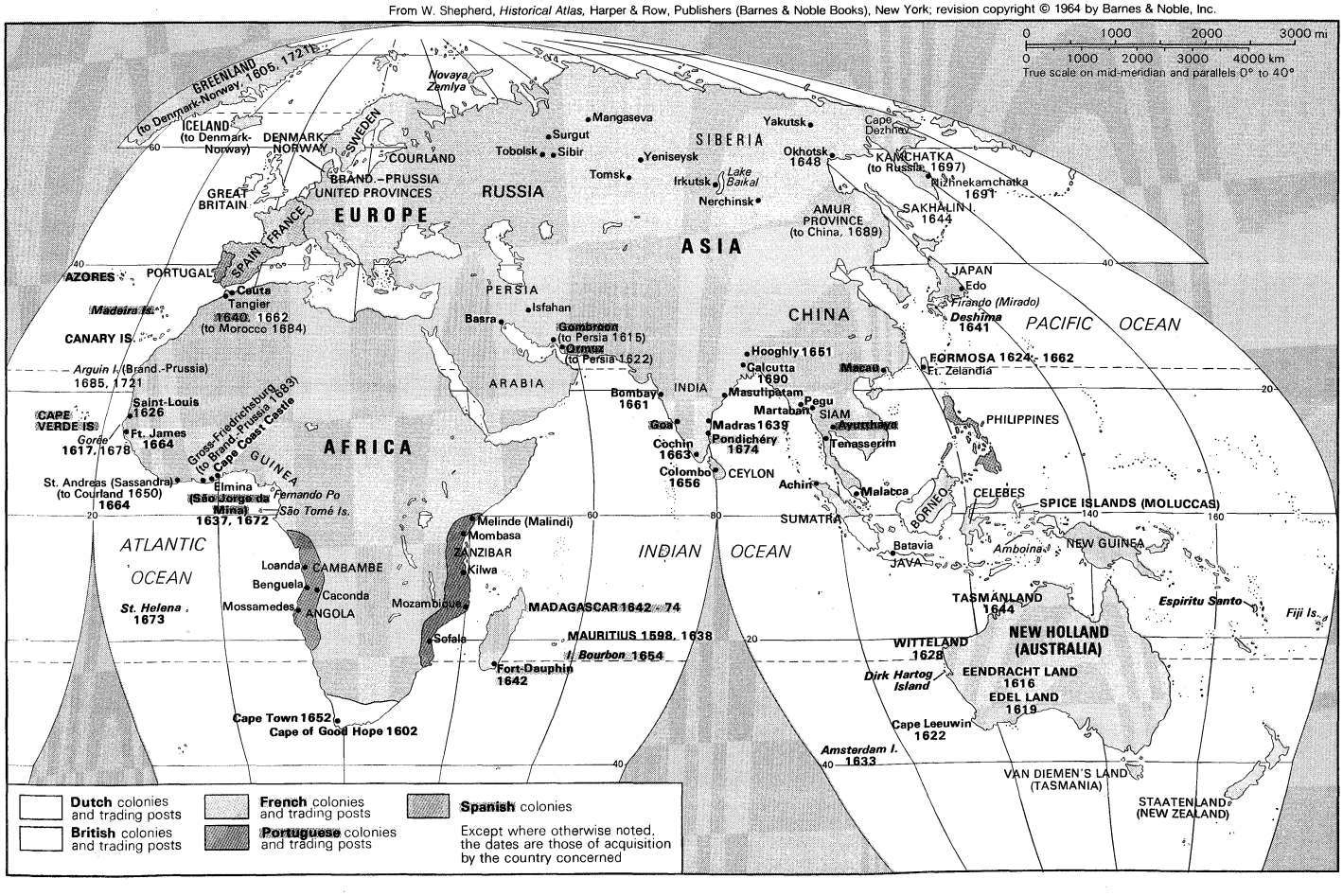
In the 18th century, with the pioneering period over, life in New France became easygoing and even pleasant, despite governmental absolutism. But the fur trade in the west drew vigorous young men from the seigneurial estates to become *coureurs de bois* (fur traders), and their loss crippled agriculture. Civil and religious authorities tried to hold settlers to farming because furs paid neither tithes nor seigneurial dues. This drainage of manpower partly explains the slow growth of New France, which, by a census of 1754, had only 55,000 whites.

The English. There is evidence that Bristol seamen reached Newfoundland before 1497, but John Cabot's Atlantic crossing in that year is the first recorded English exploration. After the death of Henry VII in 1509, England lost interest in discovery and did not resume it until 1553 and the formation of the Muscovy Company, which tried to find a Northeast Passage to Asia, discovered the island of Novaya Zemlya, and opened a small trade with Russia. The English also searched for a Northwest Passage, and Martin Frobisher sailed to Greenland, Baffin Island, and the adjacent mainland.

English ascendancy in India. Francis Drake and others raided the Spanish Main, and Drake and Thomas Cavendish sailed around the world. The defeat of Philip II's Armada in 1588, though less disastrous to Spain's seapower than commonly assumed, contributed to opening the way for English colonization of America. Interest in the Orient at first proved greater, however, and, in 1600, London merchants formed an East India Company. It could not compete with the rival Dutch company in the region of largest profits—the East Indies—so it transferred its emphasis to the Indian subcontinent. The English acquired Masulipatam in 1611 and Madras in 1639, having meanwhile destroyed Portuguese Hormuz in 1622. Charles II obtained Bombay in 1661, as part of his Portuguese queen's marriage dowry, and awarded it to the company.

Collapse of the Mughal Empire after 1707 led ultimately to armed conflict between the British and French companies for increased trade and influence. Dupleix had won

From W. Shepherd, *Historical Atlas,* Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.



the upper hand for France by 1748; but in the ensuing Seven Years' War (1756–63), fought between the major European powers in various parts of the world, the British company gained ascendancy in India, thanks largely to the ability of Robert Clive, and held it thereafter. Pondichéry surrendered; and, though France recovered this post by the ensuing Treaty of Paris (1763), French power in India had shrunk almost to nothing, while the British company's was now rivalled only by that of the native Marāthā confederacy.

Company profits from India came first from the familiar spices, but after 1660, Indian textiles outstripped these in importance. Cheap cloths, mainly cottons, found a mass market among the English poorer classes, though dainty fabrics for the wealthy also paid well. Imports of calicoes (inexpensive cotton fabrics from Calicut) to England grew so large that in 1721 Parliament passed the Calico Act to protect English manufacturers, forbidding the use of calico in England for apparel or for domestic purposes (repeal of the act in 1774 coincided with inventions of mechanical devices that made possible English cloth production in successful competition with Eastern fabrics).

England's American colonies. The English West Indies for many years exceeded North America in economic importance. The Lesser Antilles, earlier passed over by Spain in favour of the larger islands, lay open to any colonizer, though their ferocious Carib inhabitants sometimes gave trouble. The Leeward Islands of Antigua, St. Kitts, Nevis, and Barbados, as well as the Bermudas, were settled by Englishmen between 1609 and 1632. Barbados, at first the most important, owed its prosperity to the introduction of sugar culture about 1637. The size of landholdings increased in all the islands, and the white populations accordingly diminished as slavery came to furnish most of the raw labour. When an expedition sent by Oliver Cromwell took Spanish Jamaica in 1655, that island became the English West Indian centre. Settlement of Belize (later British Honduras) by buccaneers and log cutters began in 1636, although more than a century elapsed before Spain acknowledged that the English indeed had the right to be there.

Settlements in the West Indies

The English islanders, to the envy of their Dutch and French neighbours, enjoyed such constitutional privileges as the right to elect semipopular assemblies. Barbados once hoped to have two representatives in Parliament, and some Barbadians, during the English (Glorious) Revolution (1688–89), thought of making their island an independent state, but nothing came of this.

The original English mainland colonies—Virginia (founded 1607), Plymouth (1620), and Massachusetts Bay (1630)—were founded by joint-stock companies. The later New England settlements—New Hampshire, New Haven, Connecticut, and Rhode Island—began as offshoots of Massachusetts, which acquired jurisdiction over the Maine territory. The New England colonies were first peopled partly by religious dissenters, but except for the separatist Plymouth Pilgrims they did not formally secede from the Church of England for the time being.

Proprietary colonies, under individual entrepreneurs, began with Maryland, founded in 1634 under the Catholic direction of Cecilius and Leonard Calvert. Also proprietary was Pennsylvania, which originally included Delaware, founded by the Quaker William Penn in 1682. Maryland and Pennsylvania, except for a brief royal interlude in Maryland, continued under Calvert and Penn heirs until the American Revolution; all other colonies except Connecticut and Rhode Island ultimately had royal governments. The Carolinas, after abortive attempts at colonization, were effectively founded in 1670 and became first proprietary and, later, royal colonies. Georgia, last of the 13, began in 1732, partly as a philanthropic enterprise headed by James Oglethorpe to furnish a rehabilitation home for debtors and other underprivileged Englishmen. All the mainland colonies eventually had representative assemblies, chosen by the propertied classes, to aid and often handicap their English governors.

The original settlers, predominantly English, were later supplemented by French Huguenots, Germans, and Scots-Irish, especially in western New York, Pennsylvania, and the southern colonies. New York, acquired from the United Provinces of the Netherlands and including New Jersey, continued to have some Dutch flavour long after the Dutch had become a small minority. By the French and Indian War (1754–63, the American portion of the Seven Years' War), the total population of the mainland colonies was estimated as 1,296,000 whites and 300,000 blacks, enormously in excess of the 55,000 whites inhabiting French Canada.

The only bond of union among the British colonies was their allegiance to the king, and in the wars with France (*c.* 1689–1763) it proved hard to unite them against the common enemy. All the colonies were agricultural, with New England being a region of small farms, the Middle Atlantic colonies having a larger scaled and more diversified farming, and the southern ones tending to plantations on which tobacco, rice, and indigo were raised by slaves (although slavery was legal throughout all the colonies). There was much colonial shipping, especially from New England, whose merchants and seamen traded with England, Africa, and the West Indies; Massachusetts shipbuilders had built more than 700 ships by 1675. By 1763 several towns had grown into cities, including Boston, New York City, Philadelphia, Baltimore, and Charleston, South Carolina.

Economic activities in the English colonies

Mercantilism. By the time the term mercantile system was coined in 1776 by the Scottish philosopher Adam Smith, European states had been trying for two centuries to put mercantile theory into practice. The basis of mercantilism was the notion that national wealth is measured by the amount of gold and silver a nation possesses. This seemed proven by the fact that Spain's most powerful years had occurred when it was first reaping a bullion harvest from its overseas possessions.

The mercantile theory held that colonies exist for the economic benefit of the mother country and are useless unless they help to achieve profit. The mother nation should draw raw materials from its possessions and sell them finished goods, with the balance favouring the European country. This trade should be monopolistic, with foreign intruders barred.

The Spanish fleet system. Spain acted upon the as-yet-undefined mercantile theory when, in 1565, it perfected the fleet (*flota*) system, by which all legal trade with its American colonies was restricted to two annual fleets between Seville and designated ports on the Gulf of Mexico and Caribbean. The outgoing ships bore manufactured articles; returning, their cargoes consisted partly of gold and silver bars. Though the system continued for nearly two centuries, Spain was a poor country by 1700.

French mercantilist activities. Ignoring this lesson, other European states adopted the mercantilist policy; the France of Louis XIV and Colbert is the outstanding example. Colbert, who dominated French policy for 20 years, strictly regulated the economy. He instituted protective tariffs and sponsored a monopolistic merchant marine. He regarded what few overseas possessions France then had as ultimate sources of liquid wealth, which they were poorly situated to furnish because they lacked such supplies of bullion as Spain controlled in Mexico and Peru.

The English navigation acts. England adhered to mercantilism for two centuries and, possessing a more lucrative empire than France, strove to implement the policy by a series of navigation acts. The first, passed by Oliver Cromwell's government in 1651, attempted chiefly to exclude the Dutch from England's carrying trade: goods imported from Africa, Asia, or America could be brought only in English ships, which included colonial vessels, thus giving the English North American merchant marine a substantial stimulus. After the royal Restoration in 1660, Parliament renewed and strengthened the Cromwellian measures. By then colonial American maritime competition with England had grown so severe that laws of 1663 required colonial ships carrying European goods to America to route them through English ports, where a duty had to be paid, but from lack of enforcement these soon became inoperative. In the early 18th century the English lost some of their enthusiasm for bullion alone and placed

chief emphasis on commerce and industry. The Molasses Act of 1733 was in the interest of the British West Indian sugar growers, who complained of the amount of French island molasses imported by the mainland colonies; the French planters had been buying fish, livestock, and lumber brought by North American ships and gladly exchanging their sugar products for them at low prices. Prohibition of colonial purchases of French molasses, though decreed, went largely unenforced, and New England, home of most of the carrying trade, continued prosperous.

THE OLD COLONIAL SYSTEM AND THE COMPETITION FOR EMPIRE (18TH CENTURY)

Faith in mercantilism waned during the 18th century, first because of the influence of French Physiocrats, who advocated the rule of nature, whereby trade and industry would be left to follow a natural course. François Quesnay, a physician at the court of Louis XV of France, led this school of thought, fundamentally advocating an agricultural economy and holding that productive land was the only genuine wealth, with trade and industry existing for the transfer of agricultural products.

Publication of *Wealth of Nations*

Adam Smith adopted some physiocratic ideas, but he considered labour very important and did not altogether accept land as the sole wealth. Smith's *Inquiry Into the Nature and Causes of the Wealth of Nations* (1776), appearing just as Britain was about to lose much of its older empire, established the basis of new economic thought—classical economics. This denigrated mercantilism and advocated free, or at least freer, trade and state noninterference with private enterprise. *Laisser-faire et laisser-aller* ("to let it alone and let it flow") became the slogan of this British economic school. Smith thought that regulation only reduced wealth, a view in part adopted by the British government 56 years after his death.

Slave trade. Slavery, though abundantly practiced in Africa itself and widespread in the ancient Mediterranean world, had nearly died out in medieval Europe. It was revived by the Portuguese in Prince Henry's time, beginning with the enslavement of Berbers in 1442. Portugal populated Cape Verde, Fernando Po (now Bioko), and São Tomé largely with black slaves and took many to the home country, especially to the regions south of the Tagus River.

New World black slavery began in 1502, when Gov. Nicolás de Ovando of Hispaniola imported a few evidently Spanish-born blacks from Spain. Rapid decimation of the Indian population of the Spanish West Indies created a labour shortage, ultimately remedied from Africa. The great reformer, Las Casas, advocated importation of blacks to replace the vanishing Indians, and he lived to regret having done so. The population of the Greater Antilles became largely black and mulatto; on the mainland, at least in the more populated parts, the Indians, supplemented by a growing mestizo caste that clung more tenaciously to life and seemed more suited to labour, kept African slavery somewhat confined to limited areas.

The Portuguese at first practiced Indian slavery in Brazil and continued to employ it partially until 1755. It was gradually replaced by the African variety, beginning prominently in the 17th century and coinciding with the rapid rise of Brazilian sugar culture.

As the English, French, Dutch, and, to a lesser extent, the Danes colonized the smaller West Indian islands, these became plantation settlements, largely cultivated by blacks. Before the latter arrived in great numbers, the bulk of manual labour, especially in the English islands, was performed by poor whites. Some were indentured, or contract, servants; some were redemptioners who agreed to pay ship captains their passage fees within a stated time or be sold to bidders; others were convicts. Some were kidnapped, with the tacit approval of the English authorities, in keeping with the mercantilist policy that advocated getting rid of the unemployed and vagrants. Black slavery eventually surpassed white servitude in the West Indies.

John Hawkins commanded the first English slave-trading expedition in 1562 and sold his cargo in the Spanish Indies. English slaving, nevertheless, remained minor until the establishment of the English island colonies in the reign of James I (ruled 1603–25). A Dutch captain sailed the first cargo of black slaves to Virginia in 1619, the year in which the colony exported 20,000 pounds (9,000 kilograms) of tobacco. The restored Stuart king, Charles II, gave English slave trade to a monopolistic company, the Royal Adventurers Trading to Africa, in 1663, but the Adventurers accomplished little because of the early outbreak of war with Holland (1665). Its successor, the Royal African Company, was founded in 1672 and held the English monopoly until 1698, when all Englishmen received the right to trade in slaves. The Royal African Company continued slaving until 1731, when it abandoned slaving in favour of traffic in ivory and gold dust. A new slaving company, the Merchants Trading to Africa (founded 1750), had directors in London, Liverpool, and Bristol, with Bristol furnishing the largest quota of ships, estimated at 237 in 1755. Jamaica offered the greatest single market for slaves and is believed to have received 610,000 between 1700 and 1786. The slave trade still flourished in 1763, when about 150 ships sailed yearly from British ports to Africa with capacity for nearly 40,000 slaves.

First blacks in English North America

There was no well-organized opposition to the slave trade before 1800, although some individuals and ephemeral societies condemned it. The Spanish church saw the importation of blacks as an opportunity for converting them. The English religionist George Fox, founder of Quakerism (founded in the 1650s), accepted the fact that his followers had bought slaves in Barbados, but he urged kind treatment. The English novelist and political pamphleteer Daniel Defoe later denounced the traffic but seemingly regarded slavery itself as inevitable. The English and Pennsylvania Quakers passed resolutions forbidding their members to engage in the trade, but their wording suggests that some were doing so; in fact, 84 of them were members of the Merchants Trading to Africa.

Those opposing the slave trade often objected on other than humanitarian grounds. Some colonials feared any further growth of the black percentage of the population. Others, who justified English slave sales to the Spanish colonies because payment was in cash, condemned the same traffic with French islanders, who paid in molasses and thus competed with nearby English sugar planters.

Colonial wars of the 18th century. From 1689 to 1763 the British and French fought four wars that were mainly European in origin but which determined the colonial situation, in some cases for two centuries. Spain entered all four, first in alliance with England and later in partnership with France, though it played a secondary role.

King William's War (War of the League of Augsburg). The war known in Europe as that of the Palatinate, League of Augsburg, or Grand Alliance, and in America as King William's War, ended indecisively, after eight years, with the Treaty of Rijswijk in 1697. No territorial changes occurred in America, and because the great Mughal emperor Aurangzeb reigned in India, very little of the conflict penetrated there.

Queen Anne's War (War of the Spanish Succession). Queen Anne's War, the American phase of the War of the Spanish Succession (1701–14), began in 1702. Childless king Charles II of Spain, dying in 1700, willed his entire possessions to Philip, grandson of Louis XIV of France. England, the United Provinces, and Austria intervened, fearing a virtual union between powerful Louis and Spain detrimental to the balance of power, and Queen Anne's War lasted until terminated by the Treaty of Utrecht in 1713. England (Great Britain after 1707) gained Gibraltar and Minorca and, in North America, acquired Newfoundland and French Acadia (renamed Nova Scotia). It also received clear title to the northern area being exploited by the Hudson's Bay Company. Bourbon prince Philip was recognized as king of Spain, but the British secured the important *asiento,* or right to supply Spanish America with slaves, for 30 years.

British gains in North America

King George's War (War of the Austrian Succession). There followed a peace almost unbroken until 1739, when, with the *asiento* about to expire and Spain unwilling to renew it, Great Britain and Spain went to war. The recent amputation of an English seaman's ear by a Spanish Caribbean coast guard caused the conflict to be named

the War of Jenkins' Ear. This merged in 1740 with the War of the Austrian Succession (called King George's War in America), between Frederick II the Great of Prussia and Maria Theresa of Austria over Silesia. France joined Spain and Prussia against Great Britain and Austria, and the war, which was terminated in 1748 by the Treaty of Aix-la-Chapelle, proved indecisive. New England colonials captured Louisbourg, the fortified French island commanding the St. Lawrence entrance, but France's progress in India counterbalanced this conquest. With the Mughal Empire now virtually extinct, the British and French East India Companies fought each other, the advantage going to the French under Dupleix, who captured Madras and nearly expelled the British. The peace treaty restored all conquests; France recovered Louisbourg, and the British regained Madras and with it another chance to become paramount in India.

The French and Indian War (the Seven Years' War). Until 1754, when the two powers resumed their conflict in the French and Indian War in America, the overseas possessions maintained a show of peace. During this prewar period the French attempted to increase their hold on the Ohio Valley and in 1754 built Fort-Duquesne at the future site of Pittsburgh. Lt. Col. George Washington with colonial forces, in 1754, and Gen. Edward Braddock with British regulars, in 1755, were defeated in attempts to dislodge them. Dupleix and his successor, Charles-Joseph Patissier, marquis de Bussy-Castelnau, increased their influence in India; but the recall of Dupleix in 1754 damaged French prospects there.

The Seven Years' War, fought in Europe by Frederick the Great of Prussia against Austria, France, and Russia, ended with his survival against overwhelming odds. His one ally, Great Britain, helped financially but could render small military assistance. Overseas, the British triumphed completely over France, aided by Spain in the last years of the war. The French at first had the upper hand in both India and America, but the turning point came after William Pitt the Elder, later earl of Chatham, assumed direction of the British war effort. In 1757 Clive won victory at Plassey over the Nawab of Bengal, an enemy of the British company; Sir Eyre Coote's victory at Wandewash in 1760, over the French governor Thomas Lally, was followed by the capture of Pondichéry.

Britain's overseas triumph over France

In America, thanks largely to the vigorous policy of Pitt, the British won repeated victories. The French forts Frontenac, Duquesne, and Carillon fell in 1758 and 1759. British generals Sir Jeffrey Amherst and James Wolfe took Louisbourg in 1758, Quebec in 1759, and Montreal in 1760, and the surrender of Montreal included that of the entire French colony. Meanwhile, Adm. Edward Hawke destroyed or immobilized the principal French line fleet at Quiberon Bay in 1759. Spanish intervention in the war in 1761 merely enabled the British to seize Havana and Manila.

The Treaty of Paris in 1763 gave Britain all North America east of the Mississippi, including Spanish Florida. France ceded the western Mississippi Valley to Spain as compensation for the loss of Florida. Besides having a clear path to domination of India in the Old World, Great Britain also gained African Senegal. In the West Indies, it returned Martinique and Guadeloupe to France for the sake of peace but remained easily second to Spain there in importance.

The first great era of colonial conflict had ended, and the British Empire, a century and a half old, had become the world's foremost overseas domain. Though exceeded in size by that of Spain, it was the wealthiest, backed by the overwhelming naval power of Great Britain. British prestige had reached a new height, greater perhaps than it would ever attain again.

(C.E.No./Ed.)

European expansion since 1763

Shift in colonial trade strategy

The global expansion of western Europe between the 1760s and the 1870s differed in several important ways from the expansionism and colonialism of previous centuries. Along with the rise of the Industrial Revolution, which economic historians generally trace to the 1760s, and the continuing spread of industrialization in the empire-building countries came a shift in the strategy of trade with the colonial world. Instead of being primarily buyers of colonial products (and frequently under strain to offer sufficient salable goods to balance the exchange), as in the past, the industrializing nations increasingly became sellers in search of markets for the growing volume of their machine-produced goods. Furthermore, over the years there occurred a decided shift in the composition of demand for goods produced in the colonial areas. Spices, sugar, and slaves became relatively less important with the advance of industrialization, concomitant with a rising demand for raw materials for industry (*e.g.*, cotton, wool, vegetable oils, jute, dyestuffs) and food for the swelling industrial areas (wheat, tea, coffee, cocoa, meat, butter).

This shift in trading patterns entailed in the long run changes in colonial policy and practice as well as in the nature of colonial acquisitions. The urgency to create markets and the incessant pressure for new materials and food were eventually reflected in colonial practices, which sought to adapt the colonial areas to the new priorities of the industrializing nations. Such adaptation involved major disruptions of existing social systems over wide areas of the globe. Before the impact of the Industrial Revolution, European activities in the rest of the world were largely confined to: (1) occupying areas that supplied precious metals, slaves, and tropical products then in large demand; (2) establishing white-settler colonies along the coast of North America; and (3) setting up trading posts and forts and applying superior military strength to achieve the transfer to European merchants of as much existing world trade as was feasible. However disruptive these changes may have been to the societies of Africa, South America, and the isolated plantation and white-settler colonies, the social systems over most of the Earth outside Europe nevertheless remained much the same as they had been for centuries (in some places for millennia). These societies, with their largely self-sufficient small communities based on subsistence agriculture and home industry, provided poor markets for the mass-produced goods flowing from the factories of the technologically advancing countries; nor were the existing social systems flexible enough to introduce and rapidly expand the commercial agriculture (and, later, mineral extraction) required to supply the food and raw material needs of the empire builders.

Adaptation of nonindustrialized regions

The adaptation of the nonindustrialized parts of the world to become more profitable adjuncts of the industrializing nations embraced, among other things: (1) overhaul of existing land and property arrangements, including the introduction of private property in land where it did not previously exist, as well as the expropriation of land for use by white settlers or for plantation agriculture; (2) creation of a labour supply for commercial agriculture and mining by means of direct forced labour and indirect measures aimed at generating a body of wage-seeking labourers; (3) spread of the use of money and exchange of commodities by imposing money payments for taxes and land rent and by inducing a decline of home industry; and (4) where the precolonial society already had a developed industry, curtailment of production and exports by native producers.

The classic illustration of this last policy is found in India. For centuries India had been an exporter of cotton goods, to such an extent that Great Britain for a long period imposed stiff tariff duties to protect its domestic manufacturers from Indian competition. Yet, by the middle of the 19th century, India was receiving one-fourth of all British exports of cotton piece goods and had lost its own export markets.

Clearly, such significant transformations could not get very far in the absence of appropriate political changes, such as the development of a sufficiently cooperative local elite, effective administrative techniques, and peace-keeping instruments that would assure social stability and environments conducive to the radical social changes imposed by a foreign power. Consistent with these purposes was the installation of new, or amendments of old, legal systems that would facilitate the operation of a money, business, and private land economy. Tying it all together

was the imposition of the culture and language of the dominant power.

New trends in colonial acquisitions

The changing nature of the relations between centres of empire and their colonies, under the impact of the unfolding Industrial Revolution, was also reflected in new trends in colonial acquisitions. While in preceding centuries colonies, trading posts, and settlements were in the main, except for South America, located along the coastline or on smaller islands, the expansions of the late 18th century and especially of the 19th century were distinguished by the spread of the colonizing powers, or of their emigrants, into the interior of continents. Such continental extensions, in general, took one of two forms, or some combination of the two: (1) the removal of the indigenous peoples by killing them off or forcing them into specially reserved areas, thus providing room for settlers from western Europe who then developed the agriculture and industry of these lands under the social system imported from the mother countries, or (2) the conquest of the indigenous peoples and the transformation of their existing societies to suit the changing needs of the more powerful militarily and technically advanced nations.

At the heart of Western expansionism was the growing disparity in technologies between those of the leading European nations and those of the rest of the world. Differences between the level of technology in Europe and some of the regions on other continents were not especially great in the early part of the 18th century. In fact, some of the crucial technical knowledge used in Europe at that time came originally from Asia. During the 18th century, however, and at an accelerating pace in the 19th and 20th centuries, the gap between the technologically advanced countries and technologically backward regions kept on increasing despite the diffusion of modern technology by the colonial powers. The most important aspect of this disparity was the technical superiority of Western armaments, for this superiority enabled the West to impose its will on the much larger colonial populations. Advances in communication and transportation, notably railroads, also became important tools for consolidating foreign rule over extensive territories. And along with the enormous technical superiority and the colonizing experience itself came important psychological instruments of minority rule by foreigners: racism and arrogance on the part of the colonizers and a resulting spirit of inferiority among the colonized.

Influences behind expansion policies

Naturally, the above description and summary telescope events that transpired over many decades and the incidence of the changes varied from territory to territory and from time to time, influenced by the special conditions in each area, by what took place in the process of conquest, by the circumstances at the time when economic exploitation of the possessions became desirable and feasible, and by the varying political considerations of the several occupying powers. Moreover, it should be emphasized that expansion policies and practices, while far from haphazard, were rarely the result of long-range and integrated planning. The drive for expansion was persistent, as were the pressures to get the greatest advantage possible out of the resulting opportunities. But the expansions arose in the midst of intense rivalry among major powers that were concerned with the distribution of power on the continent of Europe itself as well as with ownership of overseas territories. Thus, the issues of national power, national wealth, and military strength shifted more and more to the world stage as commerce and territorial acquisitions spread over larger segments of the globe. In fact, colonies were themselves often levers of military power—sources of military supplies and of military manpower and bases for navies and merchant marines. What appears, then, in tracing the concrete course of empire is an intertwining of the struggle for hegemony between competing national powers, the manoeuvring for preponderance of military strength, and the search for greatest advantage practically obtainable from the world's resources.

EUROPEAN COLONIAL ACTIVITY (1763–C. 1875)

Stages of history rarely, if ever, come in neat packages: the roots of new historical periods begin to form in earlier eras, while many aspects of an older phase linger on and help shape the new. Nonetheless, there was a convergence of developments in the early 1760s, which, despite many qualifications, delineates a new stage in European expansionism and especially in that of the most successful empire builder, Great Britain. It is not only the Industrial Revolution in Great Britain that can be traced to this period but also the consequences of England's decisive victory over France in the Seven Years' War and the beginnings of what turned out to be the second British Empire. As a result of the Treaty of Paris, France lost nearly all of its colonial empire, while Britain became, except for Spain, the largest colonial power in the world.

The second British Empire. The removal of threat from the strongest competing foreign power set the stage for Britain's conquest of India and for operations against the North American Indians to extend British settlement in Canada and westerly areas of the North American continent. In addition, the new commanding position on the seas provided an opportunity for Great Britain to probe for additional markets in Asia and Africa and to try to break the Spanish trade monopoly in South America. During this period, the scope of British world interests broadened dramatically to cover the South Pacific, the Far East, the South Atlantic, and the coast of Africa.

British naval superiority

The initial aim of this outburst of maritime activity was not so much the acquisition of extensive fresh territory as the attainment of a far-flung network of trading posts and maritime bases. The latter, it was hoped, would serve the interdependent aims of widening foreign commerce and controlling ocean shipping routes. But in the long run many of these initial bases turned out to be steppingstones to future territorial conquests. Because the indigenous populations did not always take kindly to foreign incursions into their homelands, even when the foreigners limited themselves to small enclaves, penetration of interiors was often necessary to secure base areas against attack.

Loss of the American colonies. The path of conquest and territorial growth was far from orderly. It was frequently diverted by the renewal or intensification of rivalry between, notably, England, France, Spain, and the Low Countries in colonial areas and on the European continent. The most severe blow to Great Britain's 18th-century dreams of empire, however, came from the revolt of the 13 American colonies. These contiguous colonies were at the heart of the old, or what is often referred to as the first, British Empire, which consisted primarily of Ireland, the North American colonies, and the plantation colonies of the West Indies. Ironically, the elimination of this core of the first British Empire was to a large extent influenced by the upsurge of empire building after the Seven Years' War. Great Britain harvested from its victory in that war a new expanse of territory about equal to its prewar possessions on the North American continent: French Canada, the Floridas, and the territory between the Alleghenies and the Mississippi River. The assimilation of the French Canadians, control of the Indians and settlement of the trans-Allegheny region, and the opening of new trade channels created a host of problems for the British government. Not the least of these were the burdensome costs to carry out this program on top of a huge national debt accumulated during the war. To cope with these problems, new imperial policies were adopted by the mother country: raising (for the first time) revenue from the colonies; tightening mercantile restrictions, imposing firm measures against smuggling (an important source of income for colonial merchants), and putting obstacles in the way of New England's substantial trade with the West Indies. The strains generated by these policies created or intensified the hardships of large sections of the colonial population and, in addition, disrupted the relative harmony of interests that had been built up between the mother country and important elite groups in the colonies. Two additional factors, not unrelated to the enlargement of the British Empire, fed the onset and success of the American War of Independence (1775–83): first, a lessening need for military support from the mother country once the menacing French were removed from the continent and, second, support for the American

Imperial policies in the late 18th century

Revolutionary forces from the French and Spanish, who had much to fear from the enhanced sea power and expansionism of the British.

The shock of defeat in North America was not the only problem confronting British society. Ireland—in effect, a colonial dependency—also experienced a revolutionary upsurge, giving added significance to attacks by leading British free traders against existing colonial policies and even at times against colonialism itself. But such criticism had little effect except as it may have hastened colonial administrative reforms to counteract real and potential independence movements in dependencies such as Canada and Ireland.

Conquest of India. Apart from reforms of this nature, the aftermath of American independence was a diversion of British imperial interests to other areas—the beginning of the settlement of Australia being a case in point. In terms of amount of effort and significance of results, however, the pursuit of conquest in India took first place. Starting with the assumption of control over the province of Bengal (after the Battle of Plassey, 1757) and especially after the virtual removal of French influence from the Indian Ocean, the British waged more or less continuous warfare against the Indian people and took over more and more of the interior. The Marāthās, the main source of resistance to foreign intrusion, were decisively defeated in 1803, but military resistance of one sort or another continued until the middle of the 19th century. The financing and even the military manpower for this prolonged undertaking came mainly from India itself. As British sovereignty spread, new land-revenue devices were soon instituted, which resulted in raising the revenue to finance the consolidation of power in India and the conquest of other regions, breaking up the old system of self-sufficient and self-perpetuating villages and supporting an elite whose self-interests would harmonize with British rule.

Defeat of the Marāthās

Global expansion. Except for the acquisition of additional territory in India and colonies in Sierra Leone and New South Wales, the important additions to British overseas possessions between the Seven Years' War and the end of the Napoleonic era came as prizes of victory in wars with rival European colonial powers. In 1763 the first British Empire primarily centred on North America. By 1815, despite the loss of the 13 colonies, Britain had a second empire, one that straddled the globe from Canada and the Caribbean in the Western Hemisphere around the Cape of Good Hope to India and Australia. This empire was sustained by and in turn was supported by maritime power that far exceeded that of any of Britain's European rivals.

Policy changes. The half century of global expansion is only one aspect of the transition to the second British Empire. The operations of the new empire in the longer run also reflected decisive changes in British society. The replacement of mercantile by industrial enterprise as the main source of national wealth entailed changes to make national and colonial policy more consistent with the new hierarchy of interests. The restrictive trade practices and monopolistic privileges that sustained the commercial explosion of the 16th and most of the 17th centuries—built around the slave trade, colonial plantations, and monopolistic trading companies—did not provide the most effective environment for a nation on its way to becoming the workshop of the world.

The desired restructuring of policies occurred over decades of intense political conflict: the issues were not always clearly delineated, interest groups frequently overlapped, and the balance of power between competing vested interests shifted from time to time. The issues were clearly drawn in some cases, as for example over the continuation of the British East India Company's trade monopoly. The company's export of Indian silk, muslins, and other cotton goods was seen by all who were involved in any way in the production of British textiles to be an obstacle to the development of markets for competing British manufactures. Political opposition to this monopoly was strong at the end of the 18th century, but the giant step on the road to free trade was not taken until the early decades of the 19th century (termination of the Indian trade monopoly, 1813; of the Chinese trade monopoly, 1833).

In contrast, the issues surrounding the strategic slave trade were much more complicated. The West Indies plantations relied on a steady flow of slaves from Africa. British merchants and ships profited not only from supplying these slaves but also from the slave trade with other colonies in the Western Hemisphere. The British were the leading slave traders, controlling at least half of the transatlantic slave trade by the end of the 18th century. But the influential planter and slave-trade interests had come under vigorous and unrelenting attack by religious and humanitarian leaders and organizations, who propelled the issue of abolition to the forefront of British politics around the turn of the 19th century. Historians are still unravelling the threads of conflicting arguments about the priority of causes in the final abolition of the slave trade and, later, of slavery itself, because economic as well as political issues were at play: glutted sugar markets (to which low-cost producers in competing colonies contributed) stimulated thoughts about controlling future output by limiting the supply of fresh slaves; the compensation paid to plantation owners by the British government at the time of the abolition of slavery rescued many planters from bankruptcy during a sugar crisis, with a substantial part of the compensation money being used to pay off planters' debts to London bankers. Moreover, the battle between proslavery and antislavery forces was fought in an environment in which free-trade interests were challenging established mercantilist practices and the West Indies sugar economy was in a secular decline.

Controversy over the slave trade

The British were not the first to abolish the slave trade. Denmark had ended it earlier, and the U.S. Constitution, written in 1787, had already provided for its termination in 1808. But the British Act of 1807 formally forbidding the slave trade was followed up by diplomatic and naval pressure to suppress the trade. By the 1820s Holland, Sweden, and France had also passed anti-slave-trade laws. Such laws and attempts to enforce them by no means stopped the trade, so long as there was buoyant demand for this commodity and good profit from dealing in it. Some decline in the demand for slaves did follow the final emancipation in 1833 of slaves in British possessions. On the other hand, the demand for slaves elsewhere in the Americas took on new life—*e.g.*, to work the virgin soils of Cuba and Brazil and to pick the rapidly expanding U.S. cotton crops to feed the voracious appetite of the British textile industry. Accordingly, the number of slaves shipped across the Atlantic accelerated at the same time Britain and other maritime powers outlawed this form of commerce.

Involvement in Africa. Although Britain's energetic activity to suppress the slave trade was far from effective, its diplomatic and military operations for this end led it to much greater involvement in African affairs. Additional colonies were acquired (Sierra Leone, 1808; Gambia, 1816; Gold Coast, 1821) to serve as bases for suppressing the slave trade and for stimulating substitute commerce. British naval squadrons touring the coast of Africa, stopping and inspecting suspected slavers of other nations, and forcing African tribal chiefs to sign antislavery treaties did not halt the expansion of the slave trade, but they did help Britain attain a commanding position along the west coast of Africa, which in turn contributed to the expansion of both its commercial and colonial empire.

The growth of informal empire. The transformation of the old colonial and mercantilist commercial system was completed when, in addition to the abolition of slavery and the slave trade, the Corn Laws and the Navigation Acts were repealed in the late 1840s. The repeal of the Navigation Acts acknowledged the new reality: the primacy of Britain's navy and merchant shipping. The repeal of the Corn Laws (which had protected agricultural interests) signalled the maturation of the Industrial Revolution. In the light of Britain's manufacturing supremacy, exclusivity and monopolistic trade restraints were less important than, and often detrimental to, the need for ever-expanding world markets and sources of inexpensive raw materials and food.

Repeal of the Navigation Acts and Corn Laws

With the new trade strategy, under the impetus of freer trade and technical progress, came a broadening of the concept of empire. It was found that the commercial and financial advantages of formal empire could often be derived by informal means. The development of a worldwide trade network, the growth of overseas banking, the export of capital to less advanced regions, the leading position of London's money markets—all under the shield of a powerful and mobile navy—led to Great Britain's economic preeminence and influence in many parts of the world, even in the absence of political control.

Anticolonial sentiment. The growing importance of informal empire went hand in hand with increased expressions of dissatisfaction with the formal colonial empire. The critical approach to empire came from leading statesmen, government officials in charge of colonial policy, the free traders, and the philosophic Radicals (the latter, a broad spectrum of opinion makers often labelled the Little Englanders, whose voices of dissent were most prominent in the years between 1840 and 1870). Taking the long view, however, some historians question just how much of this current of political thought was really concerned with the transformation of the British Empire into a Little England. Those who seriously considered colonial separation were for the most part thinking of the more recent white-settler colonies, such as Canada, Australia, and New Zealand, and definitely not of independence for India nor, for that matter, for Ireland. Differences of opinion among the various political factions naturally existed over the best use of limited government finance, colonial administrative tactics, how much foreign territory could in practice be controlled, and such issues as the costs of friction with the United States over Canada. Yet, while there were important differences of opinion on the choice between formal and informal empire, no important conflict arose over the desirability of continued expansion of Britain's world influence and foreign commercial activity. Indeed, during the most active period of what has been presumed to be anticolonialism, both the formal and informal empires grew substantially: new colonies were added, the territory of existing colonies was enlarged, and military campaigns were conducted to widen Britain's trading and investment area, as in the Opium Wars of the mid-19th century.

Continued imperial growth

Decline of colonial rivalry. An outstanding development in colonial and empire affairs during the period between the Napoleonic Wars and the 1870s was an evident lessening in conflict between European powers. Not that conflict disappeared entirely, but the period as a whole was one of relative calm compared with either the almost continuous wars for colonial possessions in the 18th century or the revival of intense rivalries during the latter part of the 19th and early 20th centuries. Instead of wars among colonial powers during this period, there were wars against colonized peoples and their societies, incident either to initial conquest or to the extension of territorial possessions farther into the interior. Examples are Great Britain in India, Burma, South Africa (Kaffir Wars), New Zealand (Maori Wars); France in Algeria and Indochina; the Low Countries in Indonesia; Russia in Central Asia; and the United States against the North American Indians.

Contributing to the abatement of intercolonial rivalries was the undisputable supremacy of the British Navy during these years. The increased use of steamships in the 19th century helped reinforce this supremacy: Great Britain's ample domestic coal supply and its numerous bases around the globe (already owned or newly obtained for this purpose) combined to make available needed coaling stations. Over several decades of the 19th century and until new developments toward the end of the century opened up a new age of naval rivalry, no country was in a position to challenge Britain's dominance of the seas. This may have temporarily weakened Britain's acquisitive drive: the motive of preclusive occupation of foreign territory still occurred, but it was not as pressing as at other times.

On the whole, despite the relative tranquillity and the rise of anticolonial sentiment in Britain, the era was marked by a notable wave of European expansionism. Thus, in 1800 Europe and its possessions, including former colonies, claimed title to about 55 percent of the Earth's land surface: Europe, North and South America, most of India, the Russian part of Asia, parts of the East Indies, and small sections along the coast of Africa. But much of this was merely claimed; effective control existed over a little less than 35 percent, most of which consisted of Europe itself. By 1878—that is, before the next major wave of European acquisitions began—an additional 6,500,000 square miles (16,800,000 square kilometres) were claimed; during this period, control was consolidated over the new claims and over all the territory claimed in 1800. Hence, from 1800 to 1878, actual European rule (including former colonies in North and South America) increased from 35 to 67 percent of the Earth's land surface.

Consolidation of European colonial rule

Decline of the Spanish and Portuguese empires. During the early 19th century, however, there was a conspicuous exception to the trend of colonial growth, and that was the decline of the Portuguese and Spanish empires in the Western Hemisphere. The occasion for the decolonization was provided by the Napoleonic Wars. The French occupation of the Iberian Peninsula in 1807, combined with the ensuing years of intense warfare until 1814 on that peninsula between the British and French and their respective allies, effectively isolated the colonies from their mother countries. During this isolation the long-smouldering discontents in the colonies erupted in influential nationalist movements, revolutions of independence, and civil wars. The stricken mother countries could hardly interfere with events on the South American continent, nor did they have the resources, even after the Peninsular War was over, to bring enough soldiers and armaments across the Atlantic to suppress the independence forces.

Great Britain could have intervened on behalf of Spain and Portugal, but it declined. British commerce with South America had blossomed during the Napoleonic Wars. New vistas of potentially profitable opportunities opened up in those years, in contrast with preceding decades when British penetration of Spanish colonial markets consisted largely of smuggling to get past Spain's mercantile restrictions. The British therefore now favoured independence for these colonies and had little interest in helping to reimpose colonial rule, with its accompanying limitations on British trade and investment. Support for colonial independence by the British came in several ways: merchants and financiers provided loans and supplies needed by insurrectionary governments; the Royal Navy protected the shipment of those supplies and the returning specie; and the British government made it clear to other nations that it considered South American countries independent. The British forthright position on independence, as well as the availability of the Royal Navy to support this policy, gave substance to the U.S. Monroe Doctrine (1823), which the United States had insufficient strength at that time to really enforce.

After some 15 years of uprisings and wars, Spain by 1825 no longer had any colonies in South America itself, retaining only the islands of Cuba and Puerto Rico. During the same period Brazil achieved its independence from Portugal. The advantages to the British economy made possible by the consequent opening up of the Latin-American ports were eagerly pursued, facilitated by commercial treaties signed with these young nations. The reluctance of France to recognize their new status delayed French penetration of their markets and gave an advantage to the British. In one liberated area after another, brokers and commercial agents arrived from England to ferret out business opportunities. Soon the continent was flooded with British goods, often competing with much weaker native industries. Actually, Latin America provided the largest single export market for British cotton textiles in the first half of the 19th century.

South American independence

Despite the absence of formal empire, the British were able to attain economic preeminence in South America. Spanish and Portuguese colonialism had left a heritage of disunity and conflict within regions of new nations and between nations, along with conditions that led to unstable alliances of ruling elite groups. While this combination of weaknesses militated against successful self-development,

it was fertile ground for energetic foreign entrepreneurs, especially those who had technically advanced manufacturing capacities, capital resources, international money markets, insurance and shipping facilities, plus supportive foreign policies. The early orgy of speculative loans and investments soon ended. But before long, British economic penetration entered into more lasting and self-perpetuating activities, such as promoting Latin-American exports, providing railroad equipment, constructing public works, and supplying banking networks. Thus, while the collapse of the Spanish and Portuguese empires led to the decline of colonialism in the Western Hemisphere, it also paved the way for a significant expansion of Britain's informal empire of trade, investment, and finance during the 19th century.

The emigration of European peoples. European influence around the globe increased with each new wave of emigration from Europe. Tides of settlers brought with them the Old World culture and, often, useful agricultural and industrial skills. An estimated 55,000,000 Europeans left their native lands in the 100 years after 1820, the product chiefly of two forces: (1) the push to emigrate as a result of difficulties arising from economic dislocations at home and (2) the pull of land, jobs, and recruitment activities of passenger shipping lines and agents of labour-hungry entrepreneurs in the New World. Other factors were also clearly at work, such as the search for religious freedom, escape from tyrannical governments, avoidance of military conscription, and the desire for greater upward social and economic mobility. Such motives had existed throughout the centuries, however, and they are insufficient to explain the massive population movements that characterized the 19th century. Unemployment induced by rapid technological changes in agriculture and industry was an important incentive for English emigration in the mid-1800s. The surge of German emigration at roughly the same time is largely attributable to an agricultural revolution in Germany, which nearly ruined many farmers on small holdings in southwestern Germany. Under English rule, the Irish were prevented from industrial development and were directed to an economy based on export of cereals grown on small holdings. A potato blight, followed by famine and eviction of farm tenants by landlords, gave large numbers of Irish no alternative other than emigration or starvation. These three nationalities—English, German, and Irish—composed the largest group of migrants in the 1850s. In later years Italians and Slavs contributed substantially to the population spillover. The emigrants spread throughout the world, but the bulk of the population transfer went to the Americas, Siberia, and Australasia. The population outflow, greatly facilitated by European supremacy outside Europe, helped ease the social pressures and probably abated the dangers of social upheaval in Europe itself.

Spread of European culture and technology

Advance of the U.S. frontier. The outward movement of European peoples in any substantial numbers naturally was tied in with conquest and, to a greater or lesser degree, with the displacement of indigenous populations. In the United States, where by far the largest number of European emigrants went, acquisition of space for development by white immigrants entailed activity on two fronts: competition with rival European nations and disposition of the Indians. During a large part of the 19th century, the United States remained alert to the danger of encirclement by Europeans, but in addition the search for more fertile land, pursuit of the fur trade, and desire for ports to serve commerce in the Atlantic and Pacific oceans nourished the drive to penetrate the American continent. The most pressing points of tension with European nations were eliminated during the first half of the century: purchase of the Louisiana Territory from France in 1803 gave the United States control over the heartland of the continent; settlement of the War of 1812 ended British claims south of the 49th parallel up to the Rocky Mountains; Spain's cession of the Floridas in 1819 rounded out the Atlantic coastal frontier; and Russia's (1824) and Great Britain's (1846) relinquishment of claims to the Oregon territory gave the United States its window on the Pacific. The expansion of the United States, however, was not confined to liquidating rival claims of overseas empires; it also involved taking territory from neighbouring Mexico. Settlers from the United States wrested Texas from Mexico (1836), and war against Mexico (1846–48) led to the U.S. annexation of the southwestern region between New Mexico and Utah to the Pacific Ocean.

Penetration of the American continent

Diplomatic and military victories over the European nations and Mexico were but one precondition for the transcontinental expansion of the United States. In addition, the Indian tribes sooner or later had to be rooted out to clear the new territory. At times, treaties were arranged with Indian tribes, by which vast areas were opened up for white settlement. But even where peaceful agreements had been reached, the persistent pressure of the search for land and commerce created recurrent wars with Indian tribes that were seeking to retain their homes and their land. Room for the new settlers was obtained by forced removal of natives to as yet non-white-settled land—a process that was repeated as white settlers occupied ever more territory. Massacres during wars, susceptibility to infectious European diseases, and hardships endured during forced migrations all contributed to the decline in the Indian population and the weakening of its resistance. Nevertheless, Indian wars occupied the U.S. Army's attention during most of the 19th century, ending with the eventual isolation of the surviving Indians on reservations set aside by the U.S. government.

THE NEW IMPERIALISM (C. 1875–1914)

Reemergence of colonial rivalries. Although there are sharp differences of opinion over the reasons for, and the significance of, the "new imperialism," there is little dispute that at least two developments in the late 19th and in the beginning of the 20th century signify a new departure: (1) notable speedup in colonial acquisitions; (2) an increase in the number of colonial powers.

New acquisitions. The annexations during this new phase of imperial growth differed significantly from the expansionism earlier in the 19th century. While the latter was substantial in magnitude, it was primarily devoted to the consolidation of claimed territory (by penetration of continental interiors and more effective rule over indigenous populations) and only secondarily to new acquisitions. On the other hand, the new imperialism was characterized by a burst of activity in carving up as yet independent areas: taking over almost all Africa, a good part of Asia, and many Pacific islands. This new vigour in the pursuit of colonies is reflected in the fact that the rate of new territorial acquisitions of the new imperialism was almost three times that of the earlier period. Thus, the increase in new territories claimed in the first 75 years of the 19th century averaged about 83,000 square miles (215,000 square kilometres) a year. As against this, the colonial powers added an average of about 240,000 square miles (620,000 square kilometres) a year between the late 1870s and World War I (1914–18). By the beginning of that war, the new territory claimed was for the most part fully conquered, and the main military resistance of the indigenous populations had been suppressed. Hence, in 1914, as a consequence of this new expansion and conquest on top of that of preceding centuries, the colonial powers, their colonies, and their former colonies extended over approximately 85 percent of the Earth's surface. Economic and political control by leading powers reached almost the entire globe, for, in addition to colonial rule, other means of domination were exercised in the form of spheres of influence, special commercial treaties, and the subordination that lenders often impose on debtor nations.

Increase in new territories in Africa and Asia

New colonial powers. This intensification of the drive for colonies reflected much more than a new wave of overseas activities by traditional colonial powers, including Russia. The new imperialism was distinguished particularly by the emergence of additional nations seeking slices of the colonial pie: Germany, the United States, Belgium, Italy, and, for the first time, an Asian power, Japan. Indeed, this very multiplication of colonial powers, occurring in a relatively short period, accelerated the tempo of colonial growth. Unoccupied space that could

potentially be colonized was limited. Therefore, the more nations there were seeking additional colonies at about the same time, the greater was the premium on speed. Thus, the rivalry among the colonizing nations reached new heights, which in turn strengthened the motivation for preclusive occupation of territory and for attempts to control territory useful for the military defense of existing empires against rivals.

The impact of the new upsurge of rivalry is well illustrated in the case of Great Britain. Relying on its economic preeminence in manufacturing, trade, and international finance as well as on its undisputed mastery of the seas during most of the 19th century, Great Britain could afford to relax in the search for new colonies, while concentrating on consolidation of the empire in hand and on building up an informal empire. But the challenge of new empire builders, backed up by increasing naval power, put a new priority on Britain's desire to extend its colonial empire. On the other hand, the more that potential colonial space shrank, the greater became the urge of lesser powers to remedy disparities in size of empires by redivision of the colonial world. The struggle over contested space and for redivision of empire generated an increase in wars among the colonial powers and an intensification of diplomatic manoeuvring.

Struggle over redivision of empire

Rise of new industrialized nations. Parallel with the emergence of new powers seeking a place in the colonial sun and the increasing rivalry among existing colonial powers was the rise of industrialized nations able and willing to challenge Great Britain's lead in industry, finance, and world trade. In the mid-19th century Britain's economy outdistanced by far its potential rivals. But, by the last quarter of that century, Britain was confronted by restless competitors seeking a greater share of world trade and finance; the Industrial Revolution had gained a strong foothold in these nations, which were spurred on to increasing industrialization with the spread of railroad lines and the maturation of integrated national markets.

Moreover, the major technological innovations of the late 19th and early 20th centuries improved the competitive potential of the newer industrial nations. Great Britain's advantage as the progenitor of the first Industrial Revolution diminished substantially as the newer products and sources of energy of what has been called a second Industrial Revolution began to dominate industrial activity. The late starters, having digested the first Industrial Revolution, now had a more equal footing with Great Britain: they were all starting out more or less from the same base to exploit the second Industrial Revolution. This new industrialism, notably featuring mass-produced steel, electric power and oil as sources of energy, industrial chemistry, and the internal-combustion engine, spread over western Europe, the United States, and eventually Japan.

Need for heavy capital investment

A world economy. To operate efficiently, the new industries required heavy capital investment in large-scale units. Accordingly, they encouraged the development of capital markets and banking institutions that were large and flexible enough to finance the new enterprises. The larger capital markets and industrial enterprises, in turn, helped push forward the geographic scale of operations of the industrialized nations: more capital could now be mobilized for foreign loans and investment, and the bigger businesses had the resources for the worldwide search for and development of the raw materials essential to the success and security of their investments. Not only did the new industrialism generate a voracious appetite for raw materials, but food for the swelling urban populations was now also sought in the far corners of the world. Advances in ship construction (steamships using steel hulls, twin screws, and compound engines) made feasible the inexpensive movement of bulk raw materials and food over long ocean distances. Under the pressures and opportunities of the later decades of the 19th century, more and more of the world was drawn upon as primary producers for the industrialized nations. Self-contained economic regions dissolved into a world economy, involving an international division of labour whereby the leading industrial nations made and sold manufactured products and the rest of the world supplied them with raw materials and food.

New militarism. The complex of social, political, and economic changes that accompanied the new industrialism and the vastly expanded and integrated world commerce also provided a setting for intensified commercial rivalry, the rebuilding of high tariff walls, and a revival of militarism. Of special importance militarily was the race in naval construction, which was propelled by the successful introduction and steady improvement of radically new warships that were steam driven, armour-plated, and equipped with weapons able to penetrate the new armour. Before the development of these new technologies, Britain's naval superiority was overwhelming and unchallengeable. But because Britain was now obliged in effect to build a completely new navy, other nations with adequate industrial capacities and the will to devote their resources to this purpose could challenge Britain's supremacy at sea.

The new militarism and the intensification of colonial rivalry signalled the end of the relatively peaceful conditions of the mid-19th century. The conflict over the partition of Africa, the South African War (the Boer War), the Sino-Japanese War, the Spanish-American War, and the Russo-Japanese War were among the indications that the new imperialism had opened a new era that was anything but peaceful.

The new imperialism also represented an intensification of tendencies that had originated in earlier periods. Thus, for example, the decision by the United States to go to war with Spain cannot be isolated from the long-standing interest of the United States in the Caribbean and the Pacific. The defeat of Spain and the suppression of the independence revolutions in Cuba and the Philippines gave substance to the Monroe Doctrine: the United States now became the dominant power in the Caribbean, and the door was opened for acquisition of greater influence in Latin America. Possession of the Philippines was consistent with the historic interest of the United States in the commerce of the Pacific, as it had already manifested by its long interest in Hawaii (annexed in 1898) and by an expedition by Commodore Matthew Perry to Japan (1853).

Historiographical debate. The new imperialism marked the end of vacillation over the choice of imperialist military and political policies; similar decisions to push imperialist programs to the forefront were made by the leading industrial nations over a relatively short period. This historical conjuncture requires explanation and still remains the subject of debate among historians and social scientists. The pivot of the controversy is the degree to which the new imperialism was the product of primarily economic forces and in particular whether it was a necessary attribute of the capitalist system.

Serious analysts on both sides of the argument recognize that there is a multitude of factors involved: the main protagonists of economic imperialism recognize that political, military, and ideological influences were also at work; similarly, many who dispute the economic imperialism thesis acknowledge that economic interests played a significant role. The problem, however, is one of assigning priority to causes.

Hobson's interpretation of the new imperialism

Economic imperialism. The father of the economic interpretation of the new imperialism was the British liberal economist John Atkinson Hobson. In his seminal study, *Imperialism, a Study* (first published in 1902), he pointed to the role of such drives as patriotism, philanthropy, and the spirit of adventure in advancing the imperialist cause. As he saw it, however, the critical question was why the energy of these active agents takes the particular form of imperialist expansion. Hobson located the answer in the financial interests of the capitalist class as "the governor of the imperial engine." Imperialist policy had to be considered irrational if viewed from the vantage point of the nation as a whole: the economic benefits derived were far less than the costs of wars and armaments; and needed social reforms were shunted aside in the excitement of imperial adventure. But it was rational, indeed, in the eyes of the minority of financial interest groups. The reason for this, in Hobson's view, was the persistent congestion of capital in manufacturing. The pressure of capital needing investment outlets arose in part from a maldistribution of

income: low mass consuming power blocks the absorption of goods and capital inside the country. Moreover, the practices of the larger firms, especially those operating in trusts and combines, foster restrictions on output, thus avoiding the risks and waste of overproduction. Because of this, the large firms are faced with limited opportunities to invest in expanding domestic production. The result of both the maldistribution of income and monopolistic behaviour is a need to open up new markets and new investment opportunities in foreign countries.

Hobson's study covered a broader spectrum than the analysis of what he called its economic taproot. It also examined the associated features of the new imperialism, such as political changes, racial attitudes, and nationalism. The book as a whole made a strong impression on, and greatly influenced, Marxist thinkers who were becoming more involved with the struggle against imperialism. The most influential of the Marxist studies was a small book published by Lenin in 1917, *Imperialism, the Highest Stage of Capitalism.* Despite many similarities, at bottom there is a wide gulf between Hobson's and Lenin's frameworks of analysis and also between their respective conclusions. While Hobson saw the new imperialism serving the interests of certain capitalist groups, he believed that imperialism could be eliminated by social reforms while maintaining the capitalist system. This would require restricting the profits of those classes whose interests were closely tied to imperialism and attaining a more equitable distribution of income so that consumers would be able to buy up a nation's production. Lenin, on the other hand, saw imperialism as being so closely integrated with the structure and normal functioning of an advanced capitalism that he believed that only the revolutionary overthrow of capitalism, with the substitution of Socialism, would rid the world of imperialism.

Marxist interpretation

Lenin placed the issues of imperialism in a context broader than the interests of a special sector of the capitalist class. According to Lenin, capitalism itself changed in the late 19th century; moreover, because this happened at pretty much the same time in several leading capitalist nations, it explains why the new phase of capitalist development came when it did. This new phase, Lenin believed, involves political and social as well as economic changes; but its economic essence is the replacement of competitive capitalism by monopoly capitalism, a more advanced stage in which finance capital, an alliance between large industrial and banking firms, dominates the economic and political life of society. Competition continues, but among a relatively small number of giants who are able to control large sectors of the national and international economy. It is this monopoly capitalism and the resulting rivalry generated among monopoly capitalist nations that foster imperialism; in turn, the processes of imperialism stimulate the further development of monopoly capital and its influence over the whole society.

The difference between Lenin's more complex paradigm and Hobson's shows up clearly in the treatment of capital export. Like Hobson, Lenin maintained that the increasing importance of capital exports is a key figure of imperialism, but he attributed the phenomenon to much more than pressure from an overabundance of capital. He also saw the acceleration of capital migration arising from the desire to obtain exclusive control over raw material sources and to get a tighter grip on foreign markets. He thus shifted the emphasis from the general problem of surplus capital, inherent in capitalism in all its stages, to the imperatives of control over raw materials and markets in the monopoly stage. With this perspective, Lenin also broadened the concept of imperialism. Because the thrust is to divide the world among monopoly interest groups, the ensuing rivalry extends to a struggle over markets in the leading capitalist nations as well as in the less advanced capitalist and colonial countries. This rivalry is intensified because of the uneven development of different capitalist nations: the latecomers aggressively seek a share of the markets and colonies controlled by those who got there first, who naturally resist such a redivision. Other forces—political, military, and ideological—are at play in shaping the contours of imperialist policy, but Lenin insisted that these influences germinate in the seedbed of monopoly capitalism.

Noneconomic imperialism. Perhaps the most systematic alternative theory of imperialism was proposed by Joseph Alois Schumpeter, one of the best known economists of the first half of the 20th century. His essay "Zur Soziologie des Imperialismus" ("The Sociology of Imperialism") was first published in Germany in the form of two articles in 1919. Although Schumpeter was probably not familiar with Lenin's *Imperialism* at the time he wrote his essay, his arguments were directed against the Marxist currents of thought of the early 20th century and in particular against the idea that imperialism grows naturally out of capitalism. Unlike other critics, however, Schumpeter accepted some of the components of the Marxist thesis, and to a certain extent he followed the Marxist tradition of looking for the influence of class forces and class interests as major levers of social change. In doing so, he in effect used the weapons of Marxist thought to rebut the essence of Marxist theory.

Schumpeter's interpretation

A survey of empires, beginning with the earliest days of written history, led Schumpeter to conclude that there are three generic characteristics of imperialism: (1) At root is a persistent tendency to war and conquest, often producing nonrational expansions that have no sound utilitarian aim. (2) These urges are not innate in man. They evolved from critical experiences when peoples and classes were molded into warriors to avoid extinction; the warrior mentality and the interests of warrior classes live on, however, and influence events even after the vital need for wars and conquests disappears. (3) The drift to war and conquest is sustained and conditioned by the domestic interests of ruling classes, often under the leadership of those individuals who have most to gain economically and socially from war. But for these factors, Schumpeter believed, imperialism would have been swept away into the dustbin of history as capitalist society ripened; for capitalism in its purest form is antithetical to imperialism: it thrives best with peace and free trade. Yet despite the innate peaceful nature of capitalism, interest groups do emerge that benefit from aggressive foreign conquests. Under monopoly capitalism the fusion of big banks and cartels creates a powerful and influential social group that pressures for exclusive control in colonies and protectorates, for the sake of higher profits.

Notwithstanding the resemblance between Schumpeter's discussion of monopoly and that of Lenin and other Marxists, a crucial difference does remain. Monopoly capitalism in Lenin's frame of reference is a natural outgrowth of the previous stage of competitive capitalism. But according to Schumpeter, it is an artificial graft on the more natural competitive capitalism, made possible by the catalytic effect of the residue from the preceding feudal society. Schumpeter argued that monopoly capitalism can only grow and prosper under the protection of high tariff walls; without that shield there would be large-scale industry but no cartels or other monopolistic arrangements. Because tariff walls are erected by political decisions, it is the state and not a natural economic process that promotes monopoly. Therefore, it is in the nature of the state—and especially those features that blend the heritage of the previous autocratic state, the old war machine, and feudal interests and ideas along with capitalist interests—that the cause of imperialism will be discovered. The particular form of imperialism in modern times is affected by capitalism, and capitalism itself is modified by the imperialist experience. In Schumpeter's analysis, however, imperialism is not an inevitable product of capitalism.

The state as the progenitor of imperialism

Quest for a general theory of imperialism. The main trend of academic thought in the Western world is to follow Schumpeter's conclusion—that modern imperialism is not a product of capitalism—without paying close attention to Schumpeter's sophisticated sociological analysis. Specialized studies have produced a variety of interpretations of the origin or reawakening of the new imperialism: for France, bolstering of national prestige after its defeat in the Franco-German War (1870–71); for Germany, Bismarck's design to stay in power when threatened by political rivals; for England, the desire for greater mili-

tary security in the Mediterranean and India. These reasons—along with other frequently mentioned contributing causes, such as the spirit of national and racial superiority and the drive for power—are still matters of controversy with respect to specific cases and to the problem of fitting them into a general theory of imperialism. For example, if it is found that a new colony was acquired for better military defense of existing colonies, the questions still remain as to why the existing colonies were acquired in the first place and why it was considered necessary to defend them rather than to give them up. Similarly, explanations in terms of the search for power still have to account for the close relationship between power and wealth, because in the real world adequate economic resources are needed for a nation to hold on to its power, let alone to increase it. Conversely, increasing a nation's wealth often requires power. As is characteristic of historical phenomena, imperialist expansion is conditioned by a nation's previous history and the particular situation preceding each expansionist move. Moreover, it is carried forth in the midst of a complex of political, military, economic, and psychological impulses. It would seem, therefore, that the attempt to arrive at a theory that explains each and every imperialist action—ranging from a semifeudal Russia to a relatively undeveloped Italy to an industrially powerful Germany—is a vain pursuit. But this does not eliminate the more important challenge of constructing a theory that will provide a meaningful interpretation of the almost simultaneous eruption of the new imperialism in a whole group of leading powers.

PENETRATION OF THE WEST IN ASIA

Russia's eastward expansion. European nations and Japan at the end of the 19th century spread their influence and control throughout the continent of Asia. Russia, because of its geographic position, was the only occupying power whose Asian conquests were overland.

Overland conquests

In that respect there is some similarity between Russia and the United States in the forcible outward push of their continental frontiers. But there is a significant difference: the United States advance displaced the indigenous population, with the remaining Indians becoming wards of the state. On the other hand, the Russian march across Asia resulted in the incorporation of alien cultures and societies as virtual colonies of the Russian Empire, while providing room for the absorption of Russian settlers.

Although the conquest of Siberia and the drive to the Pacific had been periodically absorbing Russia's military energies since the 16th century, the acquisition of additional Asian territory and the economic integration of previously acquired territory took a new turn in the 19th century. Previously, Russian influence in its occupied territory was quite limited, without marked alteration of the social and economic structure of the conquered peoples. Aside from looting and exacting tribute from subject tribes, the major objects of interest were the fur trade, increased commerce with China and in the Pacific, and land. But changes in 19th-century Russian society, especially those coming after the Crimean War (1853–56), signalled a new departure. First, Russia's resounding defeat in that war temporarily frustrated its aspirations in the Balkans and the Near East; but, because its dynastic and military ambitions were in no way diminished, its expansionist energies turned with increased vigour to its Asian frontiers. Second, the emancipation of the serfs (1861), which eased the feudal restrictions on the landless peasants, led to large waves of migration by Russians and Ukrainians—first to Siberia and later to Central Asia. Third, the surge of industrialization, foreign trade, and railway building in the post-Crimean War decades paved the way for the integration of Russian Asia, which formerly, for all practical purposes, had been composed of separate dependencies, and for a new type of subjugation for many of these areas, especially in Central Asia, in which the conquered societies were "colonized" to suit the political and economic needs of the conqueror.

This process of acquisition and consolidation in Asia spread out in four directions: Siberia, the Far East, the Caucasus, and Central Asia. This pursuit of tsarist ambitions for empire and for warm-water ports involved numerous clashes and conflicts along the way. Russian expansion was ultimately limited not by the fierce opposition of the native population, which was at times a stumbling block, but by the counterpressure of competitive empire builders, such as Great Britain and Japan. Great Britain and Russia were mutually alarmed as the distances between the expanding frontiers of Russia and India shortened. One point of conflict was finally resolved when both powers agreed on the delimitation of the northern border of Afghanistan. A second major area of conflict in Central Asia was settled by an Anglo-Russian treaty (1907) to divide Persia into two separate spheres of influence, leaving a nominally independent Persian nation.

Penetration of China

As in the case of Afghanistan and Persia, penetration of Chinese territory produced clashes with both the native government and other imperialist powers. At times China's preoccupation with its struggle against other invading powers eased the way for Russia's penetration. Thus, in 1860, when Anglo-French soldiers had entered Peking, Russia was able to wrest from China the Amur Province and special privileges in Manchuria (Northeast Provinces) south of the Amur River. With this as a stepping-stone, Russia took over the seacoast north of Korea and founded the town of Vladivostok. But, because the Vladivostok harbour is icebound for some four months of the year, the Russians began to pay more attention to getting control of the Korean coastline, where many good year-round harbours could be found. Attempts to acquire a share of Korea, as well as all of Manchuria, met with the resistance of Britain and Japan. Further thrusts into China beyond the Amur and maritime provinces were finally thwarted by defeat in 1905 in the Russo-Japanese War.

The partitioning of China. The evolution of the penetration of Asia was naturally influenced by a multiplicity of factors—economic and political conditions in the expanding nations, the strategy of the military officials of the latter nations, the problems facing colonial rulers in each locality, pressures arising from white settlers and businessmen in the colonies, as well as the constraints imposed by the always limited economic and military resources of the imperialist powers. All these elements were present to a greater or lesser extent at each stage of the forward push of the colonial frontiers by the Dutch in Indonesia, the French in Indochina (Vietnam, Laos, Cambodia), and the British in Malaya, Burma, and Borneo.

Yet, despite the variety of influences at work, three general types of penetration stand out. One of these is expansion designed to overcome resistance to foreign rule. Resistance, which assumed many forms ranging from outright rebellion to sabotage of colonial political and economic domination, was often strongest in the border areas farthest removed from the centres of colonial power. The consequent extension of military control to the border regions tended to arouse the fears and opposition of neighbouring states or tribal societies and thus led to the further extension of control. Hence, attempts to achieve military security prompted the addition of border areas and neighbouring nations to the original colony.

A second type of expansion was a response to the economic opportunities offered by exploitation of the colonial interiors. Traditional trade and the free play of market forces in Asia did not produce huge supplies of raw materials and food or the enlarged export markets sought by the industrializing colonial powers. For this, entrepreneurs and capital from abroad were needed, mines and plantations had to be organized, labour supplies mobilized, and money economies created. All these alien intrusions functioned best under the firm security of an accommodating alien law and order.

The third type of expansion was the result of rivalry among colonial powers. When possible, new territory was acquired or old possessions extended in order either to preclude occupation by rivals or to serve as buffers for military security against the expansions of nearby colonial powers. Where the crosscurrents of these rivalries prevented any one power from obtaining exclusive control, various substitute arrangements were arrived at: parts of a country were chipped off and occupied by one or

more of the powers; spheres of influence were partitioned; unequal commerical treaties were imposed—while the countries subjected to such treatment remained nominally independent.

The penetration of China is the outstanding example of this type of expansion. In the early 19th century the middle part of eastern Asia (Japan, Korea, and China), containing about half the Asian population, was still little affected by Western penetration. By the end of the century, Korea was on the way to becoming annexed by Japan, which had itself become a leading imperialist power. China remained independent politically, though it was already extensively dominated by outside powers. Undoubtedly, the intense rivalry of the foreign powers helped save China from being taken over outright (as India had been). China was pressed on all sides by competing powers anxious for its trade and territory: Russia from the north, Great Britain (via India and Burma) from the south and west, France (via Indochina) from the south, and Japan and the United States (in part, via the Philippines) from the east.

Rivalry over China

The Opium Wars. The first phase of the forceful penetration of China by western Europe came in the two Opium Wars. Great Britain had been buying increasing quantities of tea from China, but it had few products that China was interested in buying by way of exchange. A resulting steady drain of British silver to pay for the tea was eventually stopped by Great Britain's ascendancy in India. With British merchants in control of India's foreign trade and with the financing of this trade centred in London, a three-way exchange developed: the tea Britain bought in China was paid for by India's exports of opium and cotton to China. And because of a rapidly increasing demand for tea in England, British merchants actively fostered the profitable exports of opium and cotton from India.

An increasing Chinese addiction to opium fed a boom in imports of the drug and led to an unfavourable trade balance paid for by a steady loss of China's silver reserves. In light of the economic effect of the opium trade plus the physical and mental deterioration of opium users, Chinese authorities banned the opium trade. At first this posed few obstacles to British merchants, who resorted to smuggling. But enforcement of the ban became stringent toward the end of the 1830s; stores of opium were confiscated, and warehouses were closed down. British merchants had an additional and longstanding grievance because the Chinese limited all trade by foreigners to the port of Canton.

In June 1840 the British fleet arrived at the mouth of the Canton River to begin the Opium War. The Chinese capitulated in 1842 after the fleet reached the Yangtze, Shanghai fell, and Nanking was under British guns. The resulting Treaty of Nanking—the first in a series of commercial treaties China was forced to sign over the years—provided for: (1) cession of Hong Kong to the British crown; (2) the opening of five treaty ports, where the British would have residence and trade rights; (3) the right of British nationals in China who were accused of criminal acts to be tried in British courts; and (4) the limitation of duties on imports and exports to a modest rate. Other countries soon took advantage of this forcible opening of China; in a few years similar treaties were signed by China with the United States, France, and Russia.

Opening of China

The Chinese, however, tried to retain some independence by preventing foreigners from entering the interior of China. With the country's economic and social institutions still intact, markets for Western goods, such as cotton textiles and machinery, remained disappointing: the self-sufficient communities of China were not disrupted as those in India had been under direct British rule, and opium smuggling by British merchants continued as a major component of China's foreign trade. Western merchants sought further concessions to improve markets. But meanwhile China's weakness, along with the stresses induced by foreign intervention, was further intensified by an upsurge of peasant rebellions, especially the massive 14-year Taiping Rebellion (1850–64).

The Western powers took advantage of the increasing difficulties by pressing for even more favourable trade treaties, culminating in a second war against China (1856–60), this time by France and England. Characteristically, the Western powers invading China played a double role: in addition to forcing a new trade treaty, they also helped to sustain the Chinese ruling establishment by participating in the suppression of the Taiping Rebellion; they believed that a Taiping victory would result in a reformed and centralized China, more resistant to Western penetration. China's defeat in the second war with the West produced a series of treaties, signed at Tientsin with Britain, France, Russia, and the United States, which brought the Western world deeper into China's affairs. The Tientsin treaties provided, among other things, for the right of foreign nationals to travel in the interior, the right of foreign ships to trade and patrol on the Yangtze River, the opening up of more treaty ports, and additional exclusive legal jurisdiction by foreign powers over their nationals residing in China.

Foreign privileges in China. Treaties of this general nature were extended over the years to grant further privileges to foreigners. Furthermore, more and more Western nations—including Germany, Italy, Denmark, The Netherlands, Spain, Belgium, and Austria-Hungary—took advantage of the new opportunities by signing such treaties. By the beginning of the 20th century, some 90 Chinese ports had been opened to foreign control. While the Chinese government retained nominal sovereignty in these ports, de facto rule was exercised by one or more of the powers: in Shanghai, for example, Great Britain and the United States coalesced their interests to form the Shanghai International Settlement. In most of the treaty ports, China leased substantial areas of land at low rates to foreign governments. The consulates in these concessions exercised legal jurisdiction over their nationals, who thereby escaped China's laws and tax collections. The foreign settlements had their own police forces and tax systems and ran their own affairs independently of nominally sovereign China.

These settlements were not the only intrusion on China's sovereignty. In addition, the opium trade was finally legalized, customs duties were forced downward to facilitate competition of imported Western goods, foreign gunboats patrolled China's rivers, and aliens were placed on customs-collection staffs to ensure that China would pay the indemnities imposed by various treaties. In response to these indignities and amid growing antiforeign sentiment, the Chinese government attempted reforms to modernize and develop sufficient strength to resist foreign intrusions. Steps were taken to master Western science and technology, erect shipyards and arsenals, and build a more effective army and navy. The reforms, however, did not get very far: they did not tackle the roots of China's vulnerability, its social and political structure; and they were undertaken quite late, after foreign nations had already established a strong foothold. Also, it is likely that the reforms were not wholehearted because two opposing tendencies were at play: on the one hand, a wish to seek independence and, on the other hand, a basic reliance on foreign support by a weak Manchu government beset with rebellion and internal opposition.

Chinese attempts at modernization

The Open Door Policy. In any event, preliminary attempts to Westernize Chinese society from within did not deter further foreign penetration; nor did the subsequent revolution (1911) succeed in freeing China from Western domination. Toward the end of the 19th century, under the impact of the new imperialism, the spread of foreign penetration accelerated. Germany entered a vigorous bid for its sphere of influence; Japan and Russia pushed forward their territorial claims; and U.S. commercial and financial penetration of the Pacific, with naval vessels patrolling Chinese rivers, was growing rapidly. But at the same time this mounting foreign interest also inhibited the outright partition of China. Any step by one of the powers toward outright partition or sizable enlargement of its sphere of influence met with strong opposition from other powers. This led eventually to the Open Door Policy, advocated by the United States, which limited or restricted exclusive privileges of any one power vis-à-vis the others. It became generally accepted after the anti-foreign Boxer Rebellion (1900) in China. With the foreign armies that had been brought in to suppress the rebellion now stationed in

North China, the danger to the continued existence of the Chinese government and the danger of war among the imperialist powers for their share of the country seemed greater than ever. Agreement on the Open Door Policy helped to retain both a compliant native government and equal opportunity for commerce, finance, and investment by the more advanced nations.

Japanese resistance to colonization

Japan's rise as a colonial power. Japan was the only Asian country to escape colonization from the West. European nations and the United States tried to "open the door," and to some extent they succeeded; but Japan was able to shake off the kind of subjugation, informal or formal, to which the rest of Asia succumbed. Even more important, it moved onto the same road of industrialization as did Europe and the United States. And instead of being colonized it became one of the colonial powers.

Japan had traditionally sought to avoid foreign intrusion. For many years, only the Dutch and the Chinese were allowed trading depots, each having access to only one port. No other foreigners were permitted to land in Japan, though Russia, France, and England tried, but with little success. The first significant crack in Japan's trade and travel barriers was forced by the United States in an effort to guarantee and strengthen its shipping interests in the Far East. Japan's guns and ships were no match for those of Commodore Perry in his two U.S. naval expeditions to Japan (1853, 1854).

The Japanese, well aware of the implications of foreign penetration through observing what was happening to China, tried to limit Western trade to two ports. In 1858, however, Japan agreed to a full commercial treaty with the United States, followed by similar treaties with the Low Countries, Russia, France, and Britain. The treaty pattern was familiar: more ports were opened; resident foreigners were granted extraterritorial rights, as in China; import and export duties were predetermined, thus removing control that Japan might otherwise exercise over its foreign trade.

Many attempts have been made to explain why a weak Japan was not taken over as a colony or, at least, did not follow in China's footsteps. Despite the absence of a commonly accepted theory, two factors were undoubtedly crucial. On the one hand, the Western nations did not pursue their attempts to control Japan as aggressively as they did elsewhere. In Asia the interests of the more aggressively expanding powers had centred on India, China, and the immediately surrounding areas. When greater interest developed in a possible breakthrough in Japan in the 1850s and 1860s, the leading powers were occupied with other pressing affairs, such as thhe 1857 Indian mutiny, the Taiping Rebellion, the Crimean War, French intervention in Mexico, and the U.S. Civil War. International jealousy may also have played a role in deterring any one power from trying to gain exclusive control over the country. On the other hand, in Japan itself, the danger of foreign military intervention, a crisis in its traditional feudal society, the rise of commerce, and a disaffected peasantry led to an intense internal power struggle and finally to a revolutionary change in the country's society and a thoroughgoing modernization program, one that brought Japan the economic and military strength to resist foreign nations.

The opposing forces in Japan's civil war were lined up between the supporters of the ruling Tokugawa family, which headed a rigid hierarchical feudal society, and the supporters of the emperor Meiji, whose court had been isolated from any significant government role. The civil war culminated in 1868 in the overthrow of the Tokugawa government and the restoration of the rule of the Emperor. The Meiji Restoration also brought new interest groups to the centre of political power and instigated a radical redirection of Japan's economic development. The nub of the changeover was the destruction of the traditional feudal social system and the building of a political, social, and economic framework conducive to capitalist industrialization. The new state actively participated in the turnabout by various forms of grants and guarantees to enterprising industrialists and by direct investment in basic industries such as railways, shipbuilding, communications, and machinery. The concentration of resources in the industrial sector was matched by social reforms that eliminated feudal restrictions, accelerated mass education, and encouraged acquisition of skills in the use of Western technology. The ensuing industrialized economy provided the means for Japan to hold its own in modern warfare and to withstand foreign economic competition.

The Meiji Restoration

Soon Japan not only followed the Western path of internal industrialization, but it also began an outward aggression resembling that of the European nations. First came the acquisition and colonization of neighbouring islands: Ryukyu Islands (including Okinawa), the Kuril Islands, Bonin Islands, and Hokkaido. Next in Japan's expansion program was Korea, but the opposition of other powers postponed the transformation of Korea into a Japanese colony. The pursuit of influence in Korea involved Japan in war with China (1894–95), at the end of which China recognized Japan's interest in Korea and ceded to Japan Taiwan, the Pescadores, and southern Manchuria. At this point rival powers interceded to force Japan to forgo taking over the southern Manchuria peninsula. While France, Britain, and Germany were involved in seeking to frustrate Japan's imperial ambitions, the most direct clash was with Russia over Korea and Manchuria. Japan's defeat of Russia in the war of 1904–05 procured for Japan the lease of the Liaotung Peninsula, the southern part of the island of Sakhalin, and recognition of its "paramount interest" in Korea. Still, pressure by Britain and the United States kept Japan from fulfillment of its plan to possess Manchuria outright. By the early 20th century, however, Japan had, by means of economic and political penetration, attained a privileged position in that part of China, as well as colonies in Korea and Taiwan and neighbouring islands.

PARTITION OF AFRICA

By the turn of the 20th century, the map of Africa looked like a huge jigsaw puzzle, with most of the boundary lines having been drawn in a sort of game of give-and-take played in the foreign offices of the leading European powers. The division of Africa, the last continent to be so carved up, was essentially a product of the new imperialism, vividly highlighting its essential features. In this respect, the timing and the pace of the scramble for Africa are especially noteworthy. Before 1880 colonial possessions in Africa were relatively few and limited to coastal areas, with large sections of the coastline and almost all the interior still independent. By 1900 Africa was almost entirely divided into separate territories that were under the administration of European nations. The only exceptions were Liberia, generally regarded as being under the special protection of the United States; Morocco, conquered by France a few years later; Libya, later taken over by Italy; and Ethiopia.

The second feature of the new imperialism was also strongly evident. It was in Africa that Germany made its first major bid for membership in the club of colonial powers: between May 1884 and February 1885, Germany announced its claims to territory in South West Africa (now South West Africa/Namibia), Togoland, Cameroon, and part of the East African coast opposite Zanzibar. Two smaller nations, Belgium and Italy, also entered the ranks, and even Portugal and Spain once again became active in bidding for African territory. The increasing number of participants in itself sped up the race for conquest. And with the heightened rivalry came more intense concern for preclusive occupation, increased attention to military arguments for additional buffer zones, and, in a period when free trade was giving way to protective tariffs and discriminatory practices in colonies as well as at home, a growing urgency for protected overseas markets. Not only the wish but also the means were at hand for this carving up of the African pie. Repeating rifles, machine guns, and other advances in weaponry gave the small armies of the conquering nations the effective power to defeat the much larger armies of the peoples of Africa. Rapid railroad construction provided the means for military, political, and economic consolidation of continental interiors. With the new steamships, settlers and materials could be moved to Africa with greater dispatch, and bulk shipments of raw

materials and food from Africa, prohibitively costly for some products in the days of the sailing ship, became economically feasible and profitable.

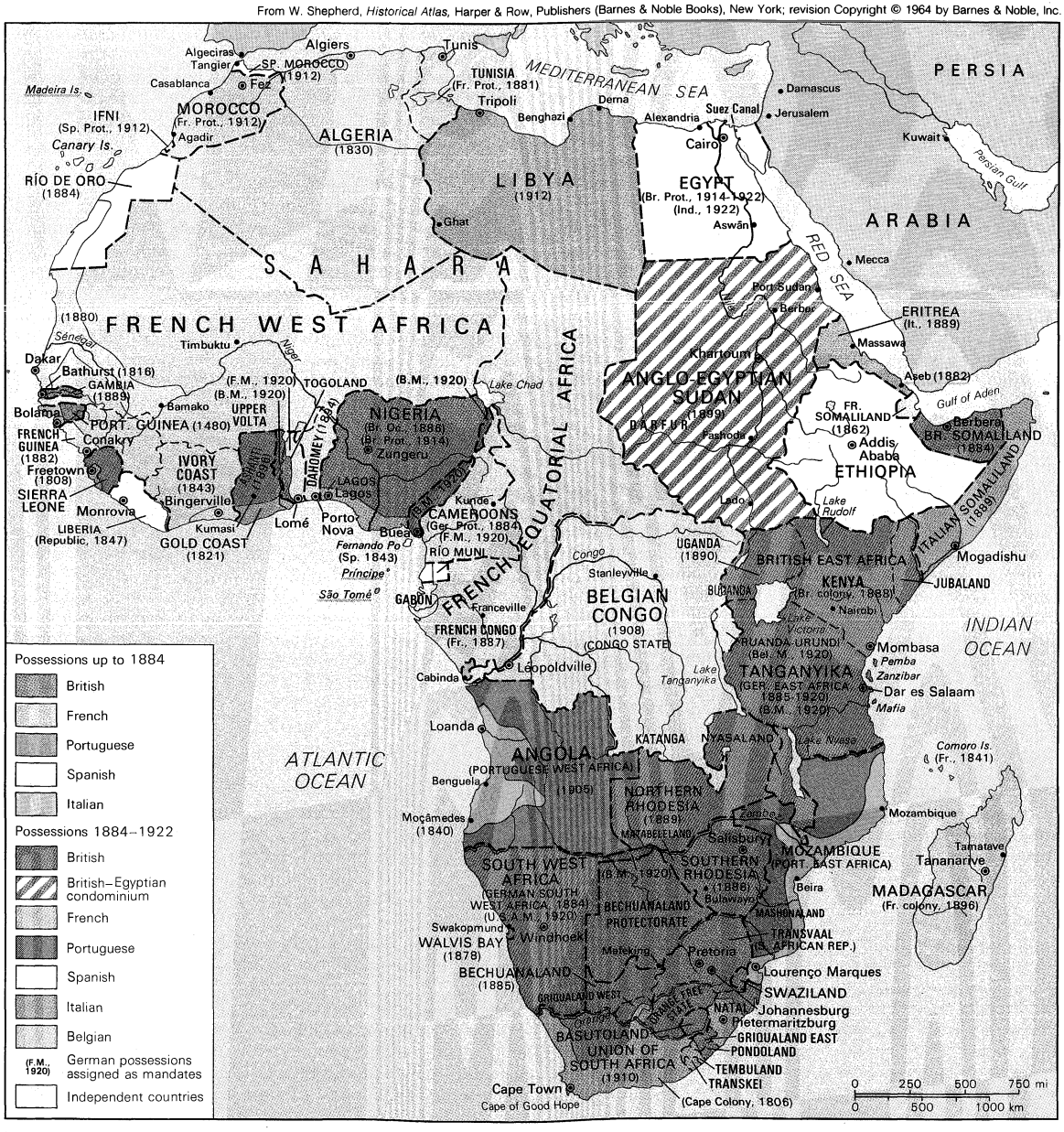
Penetration of Islāmic North Africa was complicated, on the one hand, by the struggle among European powers for control of the Mediterranean Sea and, on the other hand, by the suzerainty that the Ottoman Empire exercised to a greater or lesser extent over large sections of the region. Developments in both respects contributed to the wave of partition toward the end of the 19th century. First, Ottoman power was perceptibly waning: the military balance had tipped decisively in favour of the European nations, and Turkey was becoming increasingly dependent on loans from European centres of capital (in the late 1870s Turkey needed half of its government income just to service its foreign debt). Second, the importance of domination of the Mediterranean increased significantly after the Suez Canal was opened in 1869.

French penetration of North Africa

France was the one European nation that had established a major beachhead in Islāmic North Africa before the 1880s. At a time when Great Britain was too preoccupied to interfere, the French captured the fortress of Algiers in 1830. Frequent revolts kept the French Army busy in the Algerian interior for another 50 years before all Algeria was under full French rule. While Tunisia and Egypt had been areas of great interest to European powers during the long period of France's Algerian takeover, the penetration of these countries had been informal, confined to diplomatic and financial manoeuvres. Italy, as well as France and England, had loaned large sums to the ruling *beys* of Tunisia to help loosen that country's ties with Turkey. The inability of the *beys* to service the foreign debt in the 1870s led to the installation of debt commissioners by the lenders. Tunisia's revenues were pledged to pay the interest due on outstanding bonds; in fact, the debt charges had first call on the government's income. With this came increased pressure on the people for larger tax payments and a growing popular dissatisfaction with a government that had "sold out" to foreigners. The weakness of the ruling group, intensified by the danger of popular revolt or a military coup, opened the door further for formal occupation by one of the interested foreign powers. When Italy's actions showed that it might be preparing for outright possession, France jumped the gun by invading Tunisia in 1881 and then completed its conquest by defeating the rebellions precipitated by this occupation.

The Europeans in North Africa. The course of Egypt's loss of sovereignty resembled somewhat the same process in Tunisia: easy credit extended by Europeans, bankruptcy, increasing control by foreign-debt commissioners, mulct-

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision Copyright © 1964 by Barnes & Noble, Inc.



The European partition of Africa.

ing of the peasants to raise revenue for servicing the debt, growing independence movements, and finally military conquest by a foreign power. In Egypt, inter-imperialist rivalry, mainly between Great Britain and France, reached back to the early 19th century but was intensified under the circumstances of the new imperialism and the construction of the Suez Canal. By building the Suez Canal and financing Egypt's ruling group, France had gained a prominent position in Egypt. But Britain's interests were perhaps even more pressing because the Suez Canal was a strategic link to its empire and its other Eastern trade and colonial interests. The successful nationalist revolt headed by the Egyptian army imminently threatened in the 1880s the interests of both powers. France, occupied with war in Tunisia and with internal political problems, did not participate in the military intervention to suppress the revolt. Great Britain bombarded Alexandria in 1882, landed troops, and thus obtained control of Egypt. Unable to find a stable collaborationist government that would also pay Egypt's debts and concerned with suppressing not only the rebellion but also a powerful anti-Egyptian Mahdist revolt in the Sudan, Britain completely took over the reins of government in Egypt.

The rest of North Africa was carved up in the early 20th century. France, manoeuvring for possession of Morocco, which bordered on her Algerian colony, tried to obtain the acquiescence of the other powers by both secret and open treaties granting Italy a free hand in Libya, allotting to Spain a sphere of influence, and acknowledging Britain's paramountcy in Egypt. France had, however, overlooked Germany's ambitions, now backed by an increasingly effective army and navy. The tension created by Germany led to an international conference at Algeciras (1906), which produced a short-lived compromise, including recognition of France's paramount interest, Spanish participation in policing Morocco, and an open door for the country's economic penetration by other nations. But France's vigorous pursuit of her claims, reinforced by the occupation of Casablanca and surrounding territory, precipitated critical confrontations, which reached their peak in 1911 when French troops were suppressing a Moroccan revolt and a German cruiser appeared before Agadir in a show of force. The resulting settlements completed the European partition of North Africa: France obtained the lion's share of Morocco; in return, Germany received a large part of the French Congo; Italy was given the green light for its war with Turkey over control of Tripoli, the first step in its eventual acquisition of Libya; and Spain was enabled to extend its Río de Oro protectorate to the southern frontier of Morocco. The more or less peaceful trade-offs by the occupying powers differed sharply from the long, bitter, and expensive wars they waged against the indigenous peoples and rulers of Islāmic North Africa to solidify European rule.

The race for colonies in sub-Saharan Africa. The partition of Africa below the Sahara took place at two levels: (1) on paper—in deals made among colonial powers who were seeking colonies partly for the sake of the colonies themselves and partly as pawns in the power play of European nations struggling for world dominance—and (2) in the field—in battles of conquest against African states and tribes and in military confrontations among the rival powers themselves. This process produced, over and above the ravages of colonialism, a wasp's nest of problems that was to plague African nations long after they achieved independence. Boundary lines between colonies were often drawn arbitrarily, with little or no attention to ethnic unity, regional economic ties, tribal migratory patterns, or even natural boundaries.

Colonies in West Africa

Before the race for partition, only three European powers—France, Portugal, and Britain—had territory in tropical Africa, located mainly in West Africa. Only France had moved into the interior along the Sénégal River. The other French colonies or spheres of influence were located along the Ivory Coast and in Dahomey (now Benin) and Gabon. Portugal held on to some coastal points in Angola, Mozambique (Moçambique), and Portuguese Guinea (now Guinea-Bissau). While Great Britain had a virtual protectorate over Zanzibar in East Africa, its actual possessions were on the west coast in the Gambia, the Gold Coast, the Sierra Leone, all of them surrounded by African states that had enough organization and military strength to make the British hesitate about further expansion. Meanwhile, the ground for eventual occupation of the interior of tropical Africa was being prepared by explorers, missionaries, and traders. But such penetration remained tenuous until the construction of railroads and the arrival of steamships on navigable waterways made it feasible for European merchants to dominate the trade of the interior and for European governments to consolidate conquests.

Once conditions were ripe for the introduction of railroads and steamships in West Africa, tensions between the English and French increased as each country tried to extend its sphere of influence. As customs duties, the prime source of colonial revenue, could be evaded in uncontrolled ports, both powers began to stretch their coastal frontiers, and overlapping claims and disputes soon arose. The commercial penetration of the interior created additional rivalry and set off a chain reaction. The drive for exclusive control over interior areas intensified in response to both economic competition and the need for protection from African states resisting foreign intrusion. This drive for African possessions was intensified by the new entrants to the colonial race who felt menaced by the possibility of being completely locked out.

Perhaps the most important stimulants to the scramble for colonies south of the Sahara were the opening up of the Congo Basin by Belgium's king Leopold II and Germany's energetic annexationist activities on both the east and west coasts. As the dash for territory began to accelerate, 15 nations convened in Berlin in 1884 for the West African Conference, which, however, merely set ground rules for the ensuing intensified scramble for colonies. It also recognized the Congo Free State ruled by King Leopold, while insisting that the rivers in the Congo Basin be open to free trade. From his base in the Congo, the King subsequently took over mineral-rich Katanga, transferring both territories to Belgium in 1908.

In West Africa, Germany concentrated on consolidating its possessions of Togoland and Cameroon (Kamerun), while England and France pushed northward and eastward from their bases: England concentrated on the Niger region, the centre of its commercial activity, while France aimed at joining its possessions at Lake Chad within a grand design for an empire of contiguous territories from Algeria to the Congo. Final boundaries were arrived at after the British had defeated, among others, the Ashanti, the Fanti Confederation, the Opobo kingdom, and the Fulani; and the French won wars against the Fon kingdom, the Tuareg, the Mandingo, and other resisting tribes. The boundaries determined by conquest and agreement between the conquerors gave France the lion's share: in addition to the extension of its former coastal possessions, France acquired French West Africa and French Equatorial Africa, while Britain carved out its Nigerian colony.

In southern Africa, the intercolonial rivalries chiefly involved the British, the Portuguese, the South African Republic of the Transvaal, the British-backed Cape Colony, and the Germans. The acquisitive drive was enormously stimulated by dreams of wealth generated by the discovery of diamonds in Griqualand West and gold in Matabeleland. Encouraged by these discoveries, Cecil Rhodes (heading the British South Africa Company) and other entrepreneurs expected to find gold, copper, and diamonds in the regions surrounding the Transvaal, among them Bechuanaland, Matabeleland, Mashonaland, and Trans-Zambezia. In the ensuing struggle, which involved the conquest of the Nbele and Shona peoples, Britain obtained control over Bechuanaland and, through the British South Africa Company, over the areas later designated as the Rhodesias and Nyasaland. At the same time, Portugal moved inland to seize control over the colony of Mozambique. It was clearly the rivalries of stronger powers, especially the concern of Germany and France over the extension of British rule in southern Africa, that enabled a weak Portugal to have its way in Angola and Mozambique.

Rivalry in East Africa

The boundary lines in East Africa were arrived at largely in settlements between Britain and Germany, the two chief rivals in that region. Zanzibar and the future Tanganyika were divided in the Anglo-German treaty of 1890: Britain obtained the future Uganda and recognition of its paramount interest in Zanzibar and Pemba in exchange for ceding the strategic North Sea island of Heligoland (Helgoland) and noninterference in Germany's acquisitions in Tanganyika, Rwanda, and Urundi. Britain began to build an East African railroad to the coast, establishing the East African Protectorate (later Kenya) over the area where the railroad was to be built.

Rivalry in northeastern Africa between the French and British was based on domination of the upper end of the Nile. Italy had established itself at two ends of Ethiopia, in an area on the Red Sea that the Italians called Eritrea and in Italian Somaliland along the Indian Ocean. Italy's inland thrust led to war with Ethiopia and defeat at the hands of the Ethiopians at Adowa (1896). Ethiopia, surrounded by Italian and British armies, had turned to French advisers. The unique victory by an African state over a European army strengthened French influence in Ethiopia and enabled France to stage military expeditions from Ethiopia as well as from the Congo in order to establish footholds on the Upper Nile. The resulting race between British and French armies ended in a confrontation at Fashoda in 1898, with the British army in the stronger position. War was narrowly avoided in a settlement that completed the partition of the region: eastern Sudan was to be ruled jointly by Britain and Egypt, while France was to have the remaining Sudan from the Congo and Lake Chad to Darfur.

Germany's entrance into southern Africa through occupation and conquest of South West Africa touched off an upsurge of British colonial activity in that area, notably the separation of Basutoland (Lesotho) as a crown colony from the Cape Colony and the annexation of Zululand. As a consequence of the South African (Boer) War (1899–1902) Britain obtained sovereignty over the Transvaal and the Afrikaner Orange Free State. (Ha.Ma.)

WORLD WAR I AND THE INTERWAR PERIOD (1914–39)

Postwar redistribution of colonies. After World War I the Allied powers partitioned among themselves both the German overseas colonial holdings and the vast Arab provinces of the Ottoman Empire. They carried out this operation through the League of Nations, which awarded mandates under varying conditions. Great Britain received as mandates Iraq and Palestine (which it promptly split into Transjordan and Palestine proper); the Palestine mandate obligated Britain to respect its contradictory wartime commitments to both Jews and Arabs. France assumed a mandate over both Syria and Lebanon. In Africa the two powers divided Togo and Cameroon between them, Britain acquired Tanganyika (with a few thousand German settlers), Belgium took Rwanda–Urundi, and South Africa received German South West Africa. Italy, as compensation for not sharing in the award of mandates, obtained from Britain the Juba (Giuba) Valley on the Kenya–Somali frontier, and France eventually ceded to Italy a desert area that rounded out Libya's southern frontiers.

League mandates

The interwar years marked the apex of colonial empires throughout the world, and indirect forms of colonial penetration grew with the development of the petroleum industry. Nevertheless, most colonial systems began to show clear signs of strain and even revolt. The Russian Revolution, the Nationalist and Communist successes in China during the 1920s and '30s, the radical nationalism of Kemal Atatürk, all contributed to the rise of political movements opposed to colonialism. The very process of economic modernization, however—with the rise of factories, coordination with the world market, and mass urbanization—did more than any political or cultural factor, taken in itself, to undermine the paternal-militaristic forms of direct colonial domination.

The British Empire. Britain tended toward a decentralized and empirical type of colonial administration, in which some degree of partial decolonization could prepare the way for eventual self-rule. Realizing that direct rule over ancient civilized lands could not last indefinitely, Britain worked for a continued British presence in areas where the empire conferred self-government.

Middle East. At the outset of World War I, Britain had proclaimed a protectorate over Egypt, annulling Ottoman sovereignty; afterward, Egyptian nationalist leaders finally brought the British to recognize Egypt as an independent kingdom in 1922. In 1936–37 Egypt received control over its own economic development, and British military forces were confined to the Suez Canal area. Britain granted Iraq independence in 1932 but retained a military power base in the new kingdom. Both the world strategic balance and the British petroleum industry ruled out any possibility of a real British withdrawal from either of these Middle Eastern states.

Egypt

In Palestine the political claims of Arabs and Jews proved to be irreconcilable, and insurrection, terrorism, and occasional guerrilla warfare marked the whole period of British rule. Finally, in 1939, with war looming, the British decided to limit and eventually terminate the flow of Jewish refugees into Palestine, though not proposing to force the more than 500,000 Jewish inhabitants to live under an Arab national regime. Transjordan, detached from Palestine, became a British protectorate.

India. In India Britain faced a powerful adversary, the Indian National Congress, uniting businessmen and working classes, Hindus of high and low caste, in a common drive toward independence. The Congress never, however, succeeded in bridging the gap that separated the country's Hindu and Sikh majority from its 90,000,000 Muslims. The British met the Indian anticolonial movement half way. In 1919–23 a series of measures gave the Indians a certain degree of self-rule in a "dyarchy" in which elected Indian ministers governed together with British administrators. These constitutional reforms, however, failed to bring the princely states into line with the new trend toward self-rule. Though Mahatma Gandhi denounced the new system as a "whited sepulchre," Congress in fact began to participate in the governmental process. Under the constitution granted in 1935–37, the British maintained separate voting rolls for the Muslim minority, in order to ensure its proportional representation; in 1939 relations between Britain and the Congress Party were tense, but India was clearly headed for independence in some form.

Indian dyarchy

In 1937 the British gave a separate constitution to Burma. Ceylon (renamed Sri Lanka in 1972) had been separate and self-governing from 1931.

Africa. In British Africa decolonization progressed more slowly, but London began to accept it as an ultimate outcome. In Kenya, for example, the British government refused to grant the 20,000 European settlers in the "white highlands" any kind of direct political power over the mass of tribal blacks who constituted the colony's overwhelming majority. In British West Africa the passage from direct colonial government to self-rule by a black elite had started by 1939, there being no white settlers or Indian merchants (as there were in East Africa) to complicate matters. Only in the mining areas of Northern Rhodesia (the Copperbelt) and in Southern Rhodesia, where white farmer settlers enjoyed self-government and caste privileges over a disenfranchised black majority, did decolonization make no headway at all.

Overseas France. France, in contrast to Britain, preferred centralized and assimilative methods in an effort to integrate its colonies into a greater Overseas France. It made no progress in colonial devolution and refused even to grant independence to Syria and Lebanon. In North Africa the French energetically implanted large agrarian capitalist enterprises as well as some industries connected with the area's mineral wealth. These modern production centres and infrastructures were directed and financed by metropolitan French business and were staffed and operated by a large, politically aggressive European settler population. The Muslim majority was subordinate both politically and economically; North African peasants struggled to subsist on the margins. Overt resistance was strongest in Morocco, where a rural Muslim rebellion endangered both the French and the Spanish protectorates. Abd el-Krim, a Berber Moroccan leader who combined tradition with

Abd el-Krim

modern nationalism, waged a brilliant five-year campaign till a combined French and Spanish force finally defeated him in 1926. After 1934, resistance to France revived in Morocco, this time in the cities. In Tunisia resistance was centred in Habib Bourguiba's constitutional party; in Algeria the urban Muslim middle classes merely asked for true civil rights and integration. The French Communist Party did not move to mobilize the peasant masses in an anticolonial struggle, and, in consequence, future rebellion in the Maghrib was to be Arab nationalist and not Marxist in its leadership and doctrines.

Matters were different in French Indochina, where the growth of a modern, French-directed agricultural economy had thrown masses of peasants into debt slavery. The circumstances favoured the formation of an independence movement much influenced by both the Chinese Kuomintang (Nationalist Party) and the Chinese Communist Party; the movement in the 1930s took the form of a Communist party under the leadership of Ho Chi Minh.

French sub-Saharan Africa attracted no European settler population. The French colonial authorities promoted a shift from subsistence to market economies, and their methods, including labour conscription for public works, led to protest and questions in the French parliament. The results, guaranteed by a protective tariff linking the colonies to France, were solid but unspectacular.

Axis Powers. In the 1930s an aggressive new colonialism devoloped on the part of the Axis Powers, which developed a new colonial doctrine ("living space" in German geopolitics, the "empire" in Italian Fascist ideology, the "co-prosperity sphere" in Japan) aiming at the repartition of the world's colonial areas, justified by the supposed racial superiority, higher birth rates, and greater productivity that the Axis Powers enjoyed as against the "decadent" West. To this the Japanese added a slogan of their own, "Asia for the Asians." In fact, the three powers aimed at carving out for themselves vast, self-sufficient empires. Though intent on a new colonialism of their own, they had to use anticolonialism as a political instrument before and during World War II; in doing so, they helped in the process of world decolonization.

Fascist Italy's first colonial war was a long, bloody campaign in Cyrenaica that lasted until the early 1930s, when Italy began developing Libya as a place of settlement for Italian peasants. Then a dispute over the border between Italian Somaliland and Ethiopia (1934) gave the Italian dictator, Benito Mussolini, the opportunity to move against the African power that had routed Italian armies at Adowa. In October 1935 Italian troops from Eritrea moved into the Tigre province of northern Ethiopia, although war was never declared. Ethiopia, underequipped and feudal, could not long hold out in open combat, especially against Italian air attacks. In May 1936 Italian motorized columns reached Addis Ababa, and the Emperor went into exile. Mussolini proclaimed the Italian "empire" in East Africa. In reality, however, Ethiopian feudal chiefs continued violent resistance, even in the environs of the capital, while the Italians massacred hundreds of nobles, clergy, and commoners in an effort to repress Ethiopia by terror. In this their success was limited. The Italians built roads and kept control over all principal communication lines, but they never subdued the mountainous hinterland.

Italy in Ethiopia

The Greater East Asia Co-prosperity Sphere, Japan's new order, amounted to a self-contained empire from Manchuria to the Dutch East Indies, including China, Indochina, Thailand, and Malaya as satellite states. Japan intended to exclude both European imperialism and Communist influence from the entire Far East, while ensuring Japanese political and industrial hegemony.

The United States and the Soviet Union. During World War I the United States purchased the Virgin Islands from Denmark (1917), but it acquired no new colonies thereafter. In the 1920s the United States agreed to leave unfortified its possessions beyond Hawaii, in exchange for Japan's accepting naval limitations. The Philippines, by the Tydings–McDuffie Act of 1934, were to become independent on July 4, 1946. Until U.S.–Japanese relations began to worsen, in 1939, U.S. possessions in the Pacific counted for little in world affairs. On the other hand, the United States established or continued virtual protectorates in Cuba, Haiti, the Dominican Republic, Nicaragua, and Panama during the Harding and Coolidge administrations (1921–29), a trend reversed under Hoover and Roosevelt, particularly under the latter's Good Neighbor Policy toward Latin America.

Roosevelt's Good Neighbor Policy

The new Soviet Russian regime succeeded, after years of civil and foreign war, in regaining the Asian possessions of its tsarist predecessor. The Caucasus was repossessed step by step between 1919 and 1921; after the mountain areas and Azerbaijan were brought back under Soviet control, Armenia was partitioned between Russia and Turkey. Then Georgia, an independent parliamentary republic, was overrun by the Red Army. Russian Turkistan was subdued by 1922, and the khanates of Khiva and Bukhara were suppressed. By 1922, Outer Mongolia was also solidly linked to the Soviet state. Nevertheless, the Russian revolutionary government was ideologically opposed to colonialism, especially where it had no colonial interests that it cared to defend. In general, the Soviet authorities hesitated during the interwar period between the alternatives of backing liberation movements of "national bourgeoisies" and supporting peasant revolutionary parties.

In Central Asia the Soviet authorities followed a moderate line up to 1928, but with the advent of Stalin a new policy, consisting in purges of national leaders, increasing industrialization, and forced settlement of nomad populations, led to a great increase in the proportion of European settlers, mostly Russians and Ukrainians, to native Muslims. During the 1930s the Kazakhs declined sharply in absolute numbers as well as in ratio to the Europeans in their areas. Other Muslim nationalities, especially the Uzbeks, stemmed the Slavic tide of settlement only by virtue of their birth rates, which greatly exceeded those of the Russians and Ukrainians.

WORLD WAR II (1939–45)

Although the Axis Powers failed in their global strategy, they crippled European colonial rule in Asia.

Asia. Japan conquered its Greater East Asia Co-prosperity Sphere and arrived at the gates of India, displacing British, Dutch, and French colonial rulers as well as the Americans in Guam and the Philippines. The Japanese had to allow some margin of freedom to their satellite regimes in Burma and Indonesia in both of which preexisting local parties proved capable of creating sovereign states after the war. On August 17, 1945, Sukarno declared Indonesia independent. Indonesia had had a long history of Muslim, nationalist, and Communist agitation against the Dutch; with captured Japanese arms, Indonesia could resist reimposition of Dutch authority.

Indonesian independence

In India the Congress Party, though totally unsympathetic to the Axis, tried to take advantage of Britain's wartime extremity in order to secure immediate independence. The Muslim League supported the British administration during the war but demanded a sovereign Muslim homeland (Pakistan) as a postwar objective. By 1945 direct British rule in India was coming to an end, but the contest between Britain, the Congress Party, and the Muslim League clouded any final settlement.

Middle East. In the Middle East, Britain returned to forms of direct colonial control as Axis forces drew near, and in June–July 1941 it occupied Syria and Lebanon, under the guise of Free French administration. With Beirut and Damascus secured, the British supported Syrian and Lebanese independence from France; the two states were incorporated into the sterling area. Only U.S. and Soviet support guaranteed the independence of the two republics (1944) and their subsequent admission to the United Nations.

In Egypt, when Axis forces in 1941 and 1942 came within striking distance of Alexandria, both the king, Farouk, and groups of dissident army officers were ready to welcome them and turn against the British. In February 1942 the British minister forced the King to appoint a government willing to cooperate with the Anglo-Americans; the defeat of the Germans in the Egyptian desert later that year put Egypt firmly in the Allied camp. Nevertheless much anti-

British and anticolonial bitterness remained in Egypt, with postwar consequences.

Iran

At the outset of World War II Iran was pro-German, and in August 1941 the Soviet Union and Britain jointly occupied the country, which then became the main supply line connecting the Soviet Union with the Western Allies. In 1942, in a three-power treaty, both Britain and the Soviet Union promised to leave Iran six months after the end of the war. Notwithstanding such commitments, the Soviet Union began to build spheres of influence in northern Iran; in 1944 the Soviet Union brought pressure to bear on Iran for an oil concession.

During the final years of World War II the United States became vitally interested in the Middle East because of United States petroleum ventures in Saudi Arabia and because of strategic considerations. By the end of the war it was clear to both the Soviet Union and Britain that the United States, as a world power, would support no imposition of direct colonial controls in the postwar Middle East.

Africa. During World War II Italy lost its entire colonial domain. Ethiopia was restored as an independent empire, and the other colonies eventually came under UN jurisdiction, in the first step toward decolonization in the African continent.

DECOLONIZATION FROM 1945

In the first postwar years there were some prospects that (except in the case of the Indian subcontinent) decolonization might come gradually and on terms favourable to the continued world power positions of the western European colonial nations. After the French defeat at Dien Bien Phu (Vietnam) in 1954 and the abortive Anglo-French Suez expedition of 1956, however, decolonization took on an irresistible momentum, so that by the mid-1970s only scattered vestiges of Europe's colonial territories remained.

The reasons for this accelerated decolonization were threefold. First, the two postwar superpowers, the United States and the Soviet Union, preferred to exert their might by indirect means of penetration—ideological, economic, and military—often supplanting previous colonial rulers; both the United States and the Soviet Union took up positions opposed to colonialism. Second, the mass revolutionary movements of the colonial world fought colonial wars that were expensive and bloody. Third, the war-weary public of western Europe eventually refused any further sacrifices to maintain overseas colonies.

In general, those colonies that offered neither concentrated resources nor strategic advantages and that harboured no European settlers won easy separation from their overlords. Armed struggle against colonialism centred in a few areas, which mark the real milestones in the history of postwar decolonization.

British decolonization, 1945–56. General elections in India in 1946 strengthened the Muslim League. In subsequent negotiations, punctuated by mass violence, the Congress Party leaders finally accepted partition as preferable to civil war, and in 1947 the British evacuated the subcontinent, leaving India and a territorially divided Pakistan to contend with problems of communal strife.

Far more damaging to Britain's world position as a great power was the end of the Palestine mandate. The British would have favoured an Arab state in Palestine, tied to the British system in the Middle East, with Jews as a permanent minority. The Jewish national movement, however, succeeded in making this policy both costly and unpopular; in particular, the U.S. and Soviet governments began to see a Jewish state in Palestine as a necessary solution to the problem of Europe's surviving Jewry. All Arab spokesmen expressed intransigent opposition to any two-nation solution. Britain, isolated internationally, threw the problem into the lap of the United Nations; in November 1947 the General Assembly voted for partition. Britain, exhausted both politically and financially, decided to leave by May 15, 1948. The Jewish national movement's military branch succeeded in defeating the Palestine Arab terrorist and guerrilla bands step by step, and after British evacuation, and the declaration of Israel's independence, the Arab states in turn suffered a series of military defeats.

End of the Palestine mandate

The new Jewish state, recognized by the United States, the Soviet Union, and France, reached an uneasy armistice with the Arabs in 1949, and Britain's position in the Middle East began to crumble.

The Arab chain reaction against Britain started in Egypt, where in July 1952 a group of army officers seized power. By the end of 1954, Gamal Abdel Nasser had induced Britain to accept total withdrawal by June 1956 and set to work to undermine Britain's position in Iraq and Jordan. In June 1956 the British troops quit Suez on schedule. At that point Britain's Middle Eastern position, which depended on a chain of bases and friendly governments, was imperilled. Iran had moved close to the United States, warding off Soviet penetration and expropriating British oil holdings. Now Cyprus and the Persian Gulf oil ports remained the last outposts under British control in the Middle East. Nasser's next move was to cut the link between them. On July 26, 1956, he nationalized the Suez Canal Company, ending the last vestiges of European authority over that vital waterway and precipitating the most serious international crisis of the postwar era.

Wars in overseas France, 1945–56. The constitution of the French Fourth Republic provided for token decentralization of colonial rule, and cycles of revolt and repression marked French history for 15 years after the end of World War II. The first colonial war was in Indochina, where a power vacuum, caused by Japan's removal after wartime occupation, gave a unique opportunity to the Communist Viet Minh. When in 1946 the French Army tried to regain the colony, the Communists, proclaiming a republic, resorted to the political and military strategies of Mao Tse-tung to wear down and eventually defeat France. All chances for maintaining a semicolonial administration in Indochina ended when the Communists won the civil war in China (1949). Eventually, in 1954, when the French engaged the Communist armies in a pitched battle at Dien Bien Phu, the Communists won with the help of new heavy guns supplied by the Chinese. The Fourth Republic left Indochina under the terms of the Geneva Accords (1954), which set up two independent regimes.

Indochina War

By 1954 French North Africa was beginning to stir; guerrilla warfare occurred in both Morocco (where the French had deposed and exiled Sultan Muḥammad V) and Tunisia. On November 1, 1954, Algerian rebels began a revolt against France in which for the first time urban Muslims and Muslim peasants joined forces. In March 1956 France accorded complete independence to Morocco and Tunisia, while the army concentrated on a "revolutionary" counterinsurgent war in order to hold Algeria, where French rule had solid local support from about a million European settlers. The Muslim rebels depended on help from the Arab world, especially Egypt. Hence the French took the initiative, in October 1956, in forming an alliance with Nasser's principal adversaries, Britain and Israel, to reclaim the Suez Canal for the West and overthrow the pan-Arab regime in Cairo.

The Sinai–Suez campaign (October–November 1956). On October 29, 1956, Israel's army attacked Egypt in the Sinai Peninsula, and within 48 hours the British and French were fighting Egypt for control of the Suez area. But the Western allies found Egyptian resistance more determined than they had anticipated. Before they could turn their invasion into a real occupation, U.S. and Soviet pressure forced them to desist (November 7). The Suez campaign was thus a political disaster for the two colonial powers. The events of November 1956 showed the decline of European colonialism to be irreversible.

Algeria and French decolonization, from 1956. Between 1956 and 1958 French army commanders in Algeria, politically radicalized, tried to promote a new Franco-Muslim society in preparation for Algeria's total integration into France. Hundreds of thousands of rural Muslims were resettled under French military control, Algiers was successfully cleared of all guerrilla cells, French investments in Saharan petroleum grew, and, in a dramatic climax, a coalition of European settlers, colonial troops, and armed forces commanders in May 1958 refused further obedience to the Fourth Republic.

Charles de Gaulle, first president of the Fifth Republic,

De Gaulle and decolonization

thought that the effort of fighting colonial wars had prevented France from developing nuclear weapons and also came to realize that Algerian Muslims could not be converted to a French identity. He began to negotiate with the rebels; the negotiations culminated in a plebiscite, French evacuation, and proclamation of the independence of Muslim Algeria (July 1962). De Gaulle then proceeded to develop a nuclear striking force as the new foundation of France's status as a great power. The Fifth Republic moved rapidly toward freeing the colonies of sub-Saharan Africa, and France's colonial realm became vestigial and insular.

British decolonization after 1956. During the 15 years after the Suez disaster, Britain divested itself of most colonial holdings and abandoned most power positions in Africa and Asia. In 1958 the pro-British monarchy in Iraq fell; during the 1960s Cyprus and Malta became independent; and in 1971 Britain left the Persian Gulf. Of the imperial lifelines, only Gibraltar remains. After 1956 Britain moved rapidly to grant independence to its black African colonies. One British colony, Southern Rhodesia (now Zimbabwe), broke away unilaterally in 1965.

In Malaya the British fought a successful counterinsurgent war against a predominantly Chinese guerrilla movement and then turned over sovereignty to a federal Malaysian government (1957). In 1971 the Royal Navy left Singapore (an independent state since 1965), thus ending British presence in the Far East except at Hong Kong and (until 1983) at Brunei.

Britain's world position shrank, in effect, to membership in the North Atlantic Treaty Organization and the European Economic Community, with the postcolonial Commonwealth decreasing in importance.

Dutch, Belgian, and Portuguese decolonization. After World War II the Dutch tried to regain some of their lost control in Indonesia. The Sukarno regime held fast through three years of intermittent war, however, and the Dutch found no allies and no international support. In 1950 Indonesia became a centralized, independent republic.

The Belgian Congo

The Belgian administration in the Congo had never trained even a small number of Africans much beyond the grade-school level. When Britain and France began to divest themselves of their colonies, Belgium was in no position to impose on the Congo a schedule of its own for gradual withdrawal. The abrupt granting of independence to the Belgian Congo in the summer of 1960 led to a series of civil wars, with intervention by the UN, European business interests employing white mercenaries, and other outside forces. In 1965, Joseph Mobutu (later Mobutu Sese Seko) gained control over the central government and created an independent African state, renamed Zaire in 1971.

Portugal, in the 20th century the poorest and least developed of the western European powers, was the first nation (with Spain) to establish itself as a colonial power and the last to give up its colonial possessions. In Portuguese Africa during the authoritarian regime of António de Oliveira Salazar, the settler population had grown to about 400,000. After 1961 pan-African pressures grew, and Portugal found itself mired in a series of colonial wars, while the development of mining in Angola and Mozambique revealed hitherto unknown economic assets. In 1974 the armed forces overthrew the successors to Salazar, and in the unstable political situation it became clear that Portugal would cut its colonial ties to Africa. Portuguese Guinea (Guinea-Bissau) became independent in 1974. In June 1975 Mozambique achieved independence as a people's republic; in July 1975 São Tomé and Príncipe became an independent republic; and in November of the same year Angola, involved in a civil war between three rival liberation movements, also received sovereignty.

Conclusion. Historians will long debate the heritage of economic development, mass bitterness, and cultural cleavage that colonialism has left to the world, but the political problems of decolonization are grave and immediate. The international community is laden with minute states unable to secure either sovereignty or solvency and with large states erected without a common ethnic base. The world's postcolonial areas often have been scenes of protracted and violent conflicts: ethnic, as in Nigeria's Biafran war (1967–70); national–religious, as in the Arab–Israeli conflicts, the civil wars in Cyprus, and the continual clashes between India and Pakistan; or purely political, as in the confrontation between Communist and Nationalist regimes in the divided Korean Peninsula. Most ominously, the fading away of colonialism has created power vacuums, notably in the Indian Ocean area, into which the two superpowers have rushed. The end of colonialism has not brought with it the spread of new, neatly divided nation-states throughout the world, nor has it abated or eased rivalry between the great powers.

Post-colonial conflict

(R.A.We.)

BIBLIOGRAPHY

European exploration: HERODOTUS, *History,* trans. by J. ENOCH POWELL, 2 vol. (1949); STRABO, *Geography,* trans. by H.L. JONES, 8 vol. (1917–32); M.P. CHARLESWORTH, *Trade-Routes and Commerce of the Roman Empire,* 2nd ed. rev. (1970); PETER G. FOOTE and DAVID M. WILSON, *The Viking Achievement* (1970); GEORGE KIMBLE, *Geography in the Middle Ages* (1938, reprinted 1968); E.G.R. TAYLOR, *Tudor Geography, 1485–1583* (1930, reprinted 1968) and *Late Tudor and Early Stuart Geography, 1583–1650* (1934, reprinted 1968); EDWARD HEAWOOD, *A History of Geographical Discovery in the Seventeenth and Eighteenth Centuries* (1912, reprinted 1965); J.N.L. BAKER, *A History of Geographical Discovery and Exploration,* rev. ed. (1937), still perhaps the best single volume on the field; SIR PERCY M. SYKES, *A History of Exploration from the Earliest Times to the Present Day,* 2nd ed. (1936), a very readable account. (*On the classical period*): E.H. BUNBURY, *A History of Ancient Geography Among the Greeks and Romans,* 2nd ed., 2 vol. (1883), a standard work; J.O. THOMSON, *A History of Ancient Geography* (1948), well-documented review of geographical knowledge of the period and a discussion of the geographical theories; MAX CARY and E.H. WARMINGTON, *The Ancient Explorers* (1929), a readable account of recorded exploratory journeys. (*On the medieval period*): C.R. BEAZLEY, *The Dawn of Modern Geography,* 3 vol. (1897–1906), a standard work on geographical ideas and knowledge AD 300–1420; ARTHUR P. NEWTON (ed.), *Travel and Travellers of the Middle Ages* (1926, reprinted 1968); JOHN K. WRIGHT, *Geographical Lore of the Time of the Crusades* (1925, reprinted 1965); *The Travels of Marco Polo the Venetian*—SIR HENRY YULE edited the standard edition, but the Everyman edition is the most readily available. (*On the Age of Discovery*): A.P. NEWTON (ed.), *The Great Age of Discovery* (1932, reprinted 1969); CECIL JANE (ed.), *Select Documents Illustrating the Four Voyages of Columbus,* 2 vol. (1930–33); J.A. WILLIAMSON, *The Voyages of the Cabots . . .* (1929) and *The Cabot Voyages and Bristol Discovery Under Henry VII* (1962); F.H.H. GUILLEMARD, *Life of Ferdinand Magellan* (1890, reprinted 1971). (*On the modern period*): RICHARD HAKLUYT, *The Principall Navigations, Voiages and Discoueries of the English Nation,* 3 vol. (1598–1600)—a useful and accessible edition has been edited by JOHN MASEFIELD (1927–28); J.C. BEAGLEHOLE, *The Exploration of the Pacific,* 3rd ed. (1966); *The Journals of Captain James Cook on His Voyages of Discovery,* 3 vol. (1955–67); MARGERY PERHAM and JACK SIMMONS (eds.), *African Discovery,* 2nd ed. (1957); ERNEST SCOTT (ed.), *Australian Discovery,* 2 vol. (1929, reprinted 1966), a wide selection of passages from the journals of African and Australian explorers, with comment; CLEMENTS MARKHAM, *The Lands of Silence* (1921); ROBERT HUXLEY (ed.), *Scott's Last Expedition,* 2nd enl. ed. (1964).

European colonization (European expansion before 1763): OTTO BERKELBACH VAN DER SPRENKEL, *Die überseeische Welt und ihre Erschliessung* (1959), is a collaborative work by specialists covering all areas and subjects included here. ROMOLA and ROGER C. ANDERSON, *The Sailing-Ship,* 2nd ed. (1980), offers a concise account of sailing until the advent of steam. WILBUR CORTEZ ABBOTT, *The Expansion of Europe,* 2nd rev. ed., 2 vol. (1938), covers colonialism to 1815, with much attention to European backgrounds. JOHN H. PARRY, *The Age of Reconnaissance,* 2nd ed. (1966), is a history of discovery and conquest to 1650, offering a good scientific and maritime survey. EDGAR PRESTAGE, *The Portuguese Pioneers* (1933, reprinted 1967), is the best work in English on Portuguese voyages. CHARLES R. BOXER, *The Portuguese Seaborne Empire, 1415–1825* (1969), covers the older Portuguese empire by topics. ROGER B. MERRIMAN, *The Rise of the Spanish Empire in the Old World and the New,* 4 vol. (1918–34, reprinted 1962), follows Spain in America to the death of Philip II; LOUIS A. HARTZ (ed.), *The Founding of New Societies* (1964), presents a highly original series of essays on the colonization of Spanish America, Canada, South Africa, and the 13 colonies. *The Cambridge Economic History of Europe,* 2nd ed., vol. 4 (1967), covers the economies of the early Dutch, French, and English empires. SHEPARD B. CLOUGH and RICHARD T. RAPP, *European*

Economic History, 3rd ed. (1975), are especially good for the effects of the discoveries on Europe. GEORGE MASSELMAN, *The Cradle of Colonialism* (1963), describes the Dutch early activities in the East, providing a good European background. CHARLES R. BOXER, *The Dutch Seaborne Empire, 1600–1800* (1965), is a major work on the great age of Dutch imperialism. HERBERT I. PRIESTLEY, *France Overseas* (1938, reprinted 1966), presents a fairly good, if somewhat disjointed, account of early French overseas activity. ELI HECKSCHER, *Der Merkantilismus,* 2 vol. (1932; Eng. trans., *Mercantilism,* 2nd ed., 2 vol., 1955), is the acknowledged standard work on theoretical and historical mercantilism. ERIC WILLIAMS, *Capitalism and Slavery* (1944, reprinted 1966); and FRANK J. KLINGBERG, *The Anti-Slavery Movement in England* (1926, reprinted 1968), have chapters on the early slave trade. ALFRED T. MAHAN, *The Influence of Sea Power upon History, 1660–1783* (1890, many later editions); and LAWRENCE H. GIPSON, *The British Empire Before the American Revolution,* 15 vol. (1936–70; rev. ed., 1958–), describe the colonial wars in detail.

European expansion since 1763: DAVID K. FIELDHOUSE, *The Colonial Empires* (1966), and *Colonialism, 1870–1945* (1980), are useful general surveys of the growth and decline of empires from the 18th and 19th centuries. On the British Empire the best source is *The Cambridge History of the British Empire,* especially vol. 2, *The Growth of the New Empire, 1783–1870,* 2nd ed. (1963) and vol. 3, *The Empire Commonwealth, 1870–1919,* 2nd ed. (1967). HENRI BRUNSCHWIG, *Mythes et réalités de l'impérialisme colonial français (1871–1914)* (1960; Eng. trans., *French Colonialism, 1871–1914,* 1966), presents the case against the economic interpretation of French colonialism. A sociological study of how French colonialism operated will be found in JEAN SURET-CANALE, *L'Afrique noire: occidentale et centrale,* vol. 2, *L'Ère coloniale, 1900–1945* (1964; Eng. trans., *French Colonialism in Tropical Africa, 1900–1945,* 1971). PROSSER GIFFORD and WILLIAM R. LOUIS (eds.), *Britain and Germany in Africa* (1967), and *France and Britain in Africa* (1971), contain useful collections of essays on British, German, and French colonialism. The scramble for Africa viewed as part of Britain's striving for security in the Mediterranean and the East is forcefully argued in RONALD ROBINSON and JOHN GALLAGHER with ALICE DENNY, *Africa and the Victorians* (1961). On the growth of empire in the Far East, MICHAEL EDWARDES, *Asia in the European Age, 1498–1955* (1962), should be consulted; but see also this history as examined by an Asian in K.M. PANIKKAR, *Asia and Western Dominance,* new ed. (1959, reprinted 1969). An illuminating comparative study of colonial policies is contained in JOHN S. FURNIVALL, *Colonial Policy and Practice: A Comparative Study of Burma and Netherlands India* (1956). For a Marxist view of the impact of colonialism as related to the problems of economic development of the former colonies, see PAUL A. BARAN, *The Political Economy of Growth,* 2nd ed. (1962). U.S. expansion from the Civil War to the Spanish–American War is explored by WALTER LAFEBER in *The New Empire* (1963); and the economic aspects of U.S. expansionism are discussed by HARRY MAGDOFF in *The Age of Imperialism* (1969), and *Imperialism: From the Colonial Age to the Present* (1978). HERBERT FEIS, *Europe, the World's Banker, 1870–1914* (1930, reprinted 1964), is a useful reference work on the connection between world finance and diplomacy before World War I. A standard and detailed diplomatic history of the new imperialism is found in WILLIAM L. LANGER, *The Diplomacy of Imperialism, 1890–1902,* 2nd ed. (1950, reprinted 1965). The psychological impact of colonialism, viewed from an African perspective, is explored in FRANTZ FANON, *Les Damnés de la terre* (1961; Eng. trans., *The Wretched of the Earth,* also as *The Damned,* 1963). The case against the continuation of Western domination in the period of decolonization is found in KWAME NKRUMAH, *Neo-Colonialism: The Last Stage of Imperialism* (1965, reprinted 1981). Theories of imperialism are discussed in JOHN A. HOBSON, *Imperialism,* rev. ed. (1938, reprinted 1975); LENIN, *Imperialism, the Highest Stage of Capitalism* (1939; originally published in Russian, 1917); JOSEPH SCHUMPETER, *Imperialism and Social Classes,* ed. by PAUL M. SWEEZY (1951); ARCHIBALD P. THORNTON, *Imperialism in the Twentieth Century* (1978); and WOLFGANG J. MOMMSEN, *Imperialismustheorien,* 2nd ed. (1979; Eng. trans., *Theories of Imperialism,* (1980).

(J.B.Mi./C.E.No./Ha.Ma.)

Ancient European Religions

For roughly 20,000 years, from the Upper Paleolithic period to the beginning of the Bronze Age (*c.* 3000 BC), the continent of Europe was home to a matrifocal, pre-agrarian culture, sedentary and peaceful, extending from the eastern shores of the Black and Mediterranean seas to the Aegean and Adriatic seas. To denote this period prior to the 3rd millennium BC, by which time Indo-European invaders from the steppe region north of the Black Sea had imposed their language and their patriarchal, violent culture across the continent, archaeologists use the term "Old Europe." (The term has also been used to describe a late phase in the development of Indo-European languages.)

According to archaeological evidence, the Old Europeans worshiped a goddess represented either as a corpulent woman similar to the Paleolithic "Venus" or as a waterbird or snake-woman. The latter type, having an elongated neck and prominent buttocks, sometimes strikingly suggests the form of a phallus. An incomplete list of the goddess' companions would include the bear, the bee, the bull, the deer, the dog, the hare, the hedgehog, the hegoat, the turtle, and the toad (the last associated with the iconography of the goddess in childbirth). After the Indo-European invasions, which began in about the mid-5th millennium BC and continued for some 2,000 years, the goddess cult survived in ancient Greece and western Anatolia in the worship of such deities as Hecate, Artemis, and Kubaba. There is no consensus of interpretation among scholars regarding the iconography of the goddess, yet her absolute predominance over male representations is unmistakable. Some scholars believe that the builders of such western European megalithic monuments as Stonehenge in southwestern England, whose chronology roughly coincides with that of Old Europe, were also goddess worshipers.

Scholars have long hypothesized an underlying relationship among ancient Western languages. In 1786, in his presidential address to the Royal Asiatic Society of Bengal, the British Orientalist Sir William Jones postulated the common ancestry of Latin, Greek, and Sanskrit. The first linguist to undertake the study of this relationship was the German Franz Bopp in the 19th century. Thomas Young, the 19th-century physician and Egyptologist who helped decipher the Rosetta Stone, in 1814 coined the term "Indo-European" to encompass the ancient languages Sanskrit, Old Iranian, Hittite, Greek, and Latin, together with the Slavic, Romance, Germanic, and Celtic language groups of modern Europe. In 1819 the German philosopher Friedrich von Schlegel adopted the word "Aryan" (which properly is the name of a people who in prehistoric times settled in what is now Iran and northern India) to designate the newly discovered "race"; four years later the German Orientalist Heinrich Julius Klaproth invented the term "Indo-German," which is no more legitimate than "Indo-Slav" or "Indo-Roman," but which was adopted out of national pride. Although 19th-century philologists took quite seriously the reconstruction of the Proto-Indo-European language—which was supposed to be the common ancestor of all Indo-European languages—to the point that they conducted correspondence with one another in this artificial idiom, it has remained debatable whether or not actual linguistic unity ever existed among Indo-European peoples.

As to the existence of a Proto-Indo-European homeland, the difficult interpretation of linguistic and archaeological data has led to a proliferation of theories, most of which, however, overlap. The American archaeologist Marija Gimbutas devised the theory of the Kurgan (Turkic and Russian: "barrow," or "artificial mound") culture of seminomadic Proto-Indo-European herdsmen whose original territory encompassed the lower Volga River basin. According to Gimbutas, these patriarchal pastoralists worshiped celestial and warlike gods associated with horses, cattle, and weapons.

About 4500 BC, these people embarked on a series of broad waves of expansion, and after the second wave (*c.* 3500 BC) a secondary homeland was established in the Danube River basin, roughly coinciding with what many linguists consider to be the Proto-Indo-European homeland. Although this designation has often been restricted to the Balkan Peninsula and the northern coast of the Black Sea, it can be said to extend from southern Scandinavia to the Balkans and from the Black Sea to the Rhine River, a region that also corresponds with the diffusion of so-called corded-ware pottery. The main linguistic argument for this larger designation is based on what scholars call "macrohydronymy," *i.e.*, the names of rivers longer than 300 miles (500 kilometres) within the above-mentioned territory. Twenty-six such rivers have Indo-European names derived from roots meaning "water," "river," "marsh," and the like; and the first non-Indo-European hydronyms are Kama and Ural in the east and Liger (Loire) and Garumna (Garonne) in the west.

Thus, in the period between *c.* 3500 and *c.* 2500 BC the Indo-European languages became distributed in Europe over an area stretching from the Alps and the Rhine River in the west to the Don River in the east, and from the North and Baltic seas and the Western Dvina River in the north to the Balkan Peninsula and western Asia Minor (Anatolia) in the south.

In the late 20th century some scholars have used linguistic evidence (including hydronymy) to propose that the original homeland of the Indo-Europeans was in the Middle East and encompassed eastern Anatolia, the southern Caucasus region, and northern Mesopotamia. While this theory gained only a few adherents, it did point up the exceedingly complex and problematic nature of any such model.

If the description, or even the existence, of a common language and a common homeland of most Bronze Age Europeans (1600–1200 BC; the main exception being the Finns, a non-Indo-European, Uralic people) is open to question, all the more so does their religion elude reconstruction. Common religious themes and structures among Indo-European peoples have been emphasized since the emergence of comparative mythology in the mid-19th century. The French philologist Georges Dumézil in the mid-20th century proposed a tripartite model of Indo-European society, encompassing three groups distinguished by "social function": priests, warriors, and producers. Influenced by the French sociologist Émile Durkheim (1858–1917), who defined religion as a system of symbols encoding the rules of society, Dumézil envisioned an Indo-European religion reflecting the tripartite social order. This model, for which he found striking evidence in the major Indo-European pantheons, remained the principal focus of his work for almost 50 years. In the classic summary of his position (1958), Dumézil explicitly stated that only Indo-European societies exhibited tripartition and that its occurrence in other societies indicates Indo-European influence. Obviously, this assertion weakens his argument, but his structural scheme has opened new perspectives in the search for a common foundation among Indo-European religions. (I.P.C.)

This article treats the beliefs in and organized worship and service of gods or other supernatural powers by the various cultural groups that flourished on the continent of Europe before or concurrently with the advent of the Christian religion. For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 822, and the *Index*.

The article is divided into the following sections:

Celtic religion

The Celts, an ancient Indo-European people, reached the apogee of their influence and territorial expansion during the 4th century BC, extending across the length of Europe from Britain to Asia Minor. From the 3rd century BC onward their history is one of decline and disintegration, and with Julius Caesar's conquest of Gaul (58–51 BC) Celtic independence came to an end on the European continent. In Britain and Ireland this decline moved more slowly, but traditional culture was gradually eroded through the pressures of political subjugation; today the Celtic languages are spoken only on the western periphery of Europe, in restricted areas of Ireland, Scotland, Wales, and Brittany (in this last instance largely as a result of immigration from Britain from the 4th to the 7th century AD). It is not surprising, therefore, that the unsettled and uneven history of the Celts has affected the documentation of their culture and religion.

SOURCES

Two main types of sources provide information on Celtic religion: the sculptural monuments associated with the Celts of continental Europe and of Roman Britain, and the insular Celtic literatures that have survived in writing from medieval times. Both pose problems of interpretation. Most of the monuments, and their accompanying inscriptions, belong to the Roman period and reflect a considerable degree of syncretism between Celtic and Roman gods; even where figures and motifs appear to derive from pre-Roman tradition, they are difficult to interpret in the absence of a preserved literature on mythology. Only after the lapse of many centuries—beginning in the 7th century in Ireland, even later in Wales—was the mythological tradition consigned to writing, but by then Ireland and Wales had been Christianized and the scribes and redactors were monastic scholars. The resulting literature is abundant and varied, but it is much removed in both time and location from its epigraphic and iconographic correlatives on the

Problems of interpretation

Continent and inevitably reflects the redactors' selectivity and something of their Christian learning. Given these circumstances it is remarkable that there are so many points of agreement between the insular literatures and the continental evidence. This is particularly notable in the case of the Classical commentators from Poseidonius (*c.* 135–*c.* 51 BC) onward who recorded their own or others' observations on the Celts.

THE CELTIC GODS

The locus classicus for the Celtic gods of Gaul is the passage in Caesar's *Commentarii de bello Gallico* (52–51 BC; *The Gallic War*) in which he names five of them together with their functions. Mercury was the most honoured of all the gods and many images of him were to be found. Mercury was regarded as the inventor of all the arts, the patron of travelers and of merchants, and the most powerful god in matters of commerce and gain. After him the Gauls honoured Apollo, Mars, Jupiter, and Minerva. Of these gods they held almost the same opinions as other peoples did: Apollo drives away diseases, Minerva promotes handicrafts, Jupiter rules the heavens, and Mars controls wars.

In characteristic Roman fashion, however, Caesar does not refer to these figures by their native names but by the names of the Roman gods with which he equated them, a procedure that greatly complicates the task of identifying his Gaulish deities with their counterparts in the insular literatures. He also presents a neat schematic equation of god and function that is quite foreign to the vernacular literary testimony. Yet, given its limitations, his brief catalog is a valuable and essentially accurate witness. In comparing his account with the vernacular literatures, or even with the continental iconography, it is well to recall their disparate contexts and motivations. As has been noted, Caesar's commentary and the iconography refer to quite different stages in the history of Gaulish religion; the iconography of the Roman period belongs to an environment of profound cultural and political change, and the religion it represents may in fact have been less clearly structured than that maintained by the druids (the priestly order) in the time of Gaulish independence. On the other hand, the lack of structure is sometimes more apparent than real. It has, for instance, been noted that of the several hundred names containing a Celtic element attested in Gaul the majority occur only once, which has led some scholars to conclude that the Celtic gods and their cults were local and tribal rather than national. Supporters of this view cite Lucan's mention of a god Teutates, which they interpret as "god of the tribe" (it is thought that *teutā* meant "tribe" in Celtic). The seeming multiplicity of deity names may, however, be explained otherwise—for example, many are simply epithets applied to major deities by widely extended cults. The notion of the Celtic pantheon as merely a proliferation of local gods is contradicted by the several well-attested deities whose cults were observed virtually throughout the areas of Celtic settlement.

According to Caesar the god most honoured by the Gauls was "Mercury," and this is confirmed by numerous images and inscriptions. His Celtic name is not explicitly stated, but it is clearly implied in the place-name Lugudunon ("the fort or dwelling of the god Lugus") by which his numerous cult centres were known and from which the modern Lyon, Laon, and Loudun in France, Leiden in The Netherlands, and Legnica in Poland derive. The Irish and Welsh cognates of Lugus are Lugh and Lleu, respectively, and the traditions concerning these figures mesh neatly with those of the Gaulish god. Caesar's description of the latter as "the inventor of all the arts" might almost have been a paraphrase of Lugh's conventional epithet *sam ildánach* ("possessed of many talents"). An episode in the Irish tale of the Battle of Magh Tuiredh is a dramatic exposition of Lugh's claim to be master of all the arts and crafts, and dedicatory inscriptions in Spain and Switzerland, one of them from a guild of shoemakers, commemorate Lugus, or Lugoves, the plural perhaps referring to the god conceived in triple form. An episode in the Middle Welsh collection of tales called the *Mabinogion,* (or *Mabinogi*), seems to echo the connection with shoemaking, for it represents Lleu as working briefly as a skilled exponent of the craft. In Ireland Lugh was the youthful victor over the demonic Balar "of the venomous eye." He was the divine exemplar of sacral kingship, and his other common epithet, *lámhfhada* ("of the long arm"), perpetuates an old Indo-European metaphor for a great king extending his rule and sovereignty far afield. His proper festival, called Lughnasadh ("Festival of Lugh") in Ireland, was celebrated—and still is at several locations—in August; at least two of the early festival sites, Carmun and Tailtiu, were the reputed burial places of goddesses associated with the fertility of the earth (as was, evidently, the consort Maia—or Rosmerta ["the Provider"]—who accompanies "Mercury" on many Gaulish monuments).

The god Lugus

The Gaulish god "Mars" illustrates vividly the difficulty of equating individual Roman and Celtic deities. A famous passage in Lucan's *Bellum civile* mentions the bloody sacrifices offered to the three Celtic gods Teutates, Esus, and Taranis; of two later commentators on Lucan's text, one identifies Teutates with Mercury, the other with Mars. The probable explanation of this apparent confusion, which is paralleled elsewhere, is that the Celtic gods are not rigidly compartmentalized in terms of function. Thus "Mercury" as the god of sovereignty may function as a warrior, while "Mars" may function as protector of the tribe, so that either one may plausibly be equated with Teutates.

The problem of identification is still more pronounced in the case of the Gaulish "Apollo," for some of his 15 or more epithets may refer to separate deities. The solar connotations of Belenus (from Celtic: *bel,* "shining" or "brilliant") would have supported the identification with the Greco-Roman Apollo. Several of his epithets, such as Grannus and Borvo (which are associated etymologically with the notions of "boiling" and "heat," respectively), connect him with healing and especially with the therapeutic powers of thermal and other springs, an area of religious belief that retained much of its ancient vigour in Celtic lands throughout the Middle Ages and even to the present time. Maponos ("Divine Son" or "Divine Youth") is attested in Gaul but occurs mainly in northern Britain. He appears in medieval Welsh literature as Mabon, son of Modron (that is, of Matrona, "Divine Mother"), and he evidently figured in a myth of the infant god carried off from his mother when three nights old. His name survives in Arthurian romance under the forms Mabon, Mabuz, and Mabonagrain. His Irish equivalent was Mac ind Óg ("Young Son" or "Young Lad"), known also as Oenghus, who dwelt in Bruigh na Bóinne, the great Neolithic, and therefore pre-Celtic, passage grave of Newgrange (or Newgrange House). He was the son of Dagda (or Daghda), chief god of the Irish, and of Boann, the personified sacred river of Irish tradition. In the literature the Divine Son tends to figure in the role of trickster and lover.

The "Divine Son"

There are dedications to "Minerva" in Britain and throughout the Celtic areas of the Continent. At Bath she was identified with the goddess Sulis, whose cult there centred on the thermal springs. Through the plural form Suleviae, found at Bath and elsewhere, she is also related to the numerous and important mother goddesses—who often occur in duplicate or, more commonly, triadic form. Her nearest equivalent in insular tradition is the Irish goddess Brighid, daughter of the chief god, Dagda. Like Minerva she was concerned with healing and craftsmanship, but she was also the patron of poetry and traditional learning. Her name is cognate with that of Brigantī, Latin Brigantia, tutelary goddess of the Brigantes of Britain, and there is some onomastic evidence that her cult was known on the Continent, whence the Brigantes had migrated.

The Gaulish Sucellos (or Sucellus), possibly meaning "the Good Striker," appears on a number of reliefs and statuettes with a mallet as his attribute. He has been equated with the Irish Dagda, "the Good God," also called Eochaidh Ollathair ("Eochaidh the Great Father"), whose attributes are his club and his caldron of plenty. But, whereas Ireland had its god of the sea, Manannán mac Lir ("Manannán, son of the Ocean"), and a more shadowy predecessor called Tethra, there is no clear evidence for a Gaulish sea-god, perhaps because the original central European homeland of the Celts had been landlocked.

The "Great Father"

The insular literatures show that certain deities were associated with particular crafts. Caesar makes no mention of a Gaulish Vulcan, though insular sources reveal that there was one and that he enjoyed high status. His name in Irish, Goibhniu, and Welsh, Gofannon, derived from the Celtic word for smith. The weapons that Goibhniu forged with his fellow craft gods, the wright Luchta and the metalworker Creidhne, were unerringly accurate and lethal. He was also known for his power of healing, and as Gobbán the Wright, a popular or hypocoristic form of his name, he was renowned as a wondrous builder. Medieval Welsh also mentions Amaethon, evidently a god of agriculture, of whom little is known.

GODDESSES AND DIVINE CONSORTS

One notable feature of Celtic sculpture is the frequent conjunction of male deity and female consort, such as "Mercury" and Rosmerta, or Sucellos and Nantosvelta. Essentially these reflect the coupling of the protecting god of tribe or nation with the mother-goddess who ensured the fertility of the land. It is in fact impossible to distinguish clearly between the individual goddesses and these mother-goddesses, *matres* or *matronae,* who figure so frequently in Celtic iconography, often, as in Irish tradition, in triadic form. Both types of goddesses are concerned with fertility and with the seasonal cycle of nature, and, on the evidence of insular tradition, both drew much of their power from the old concept of a great goddess who, like the Indian Aditi, was mother of all the gods. Welsh and Irish tradition also bring out the multifaceted character of the goddess, who in her various epiphanies or avatars assumes quite different and sometimes wholly contrasting forms and personalities. She may be the embodiment of sovereignty, youthful and beautiful in union with her rightful king, or aged and hideously ugly when lacking a fitting mate. She may be the spirit of war, like the fearsome Morrígan or the Badhbh Chatha ("Raven of Battle"), whose name is attested in its Gaulish form, Cathubodua, in Haute-Savoie, or the lovely otherworld visitor who invites the chosen hero to accompany her to the land of eternal youth. As the life-giving force she is often identified with rivers, such as the Seine (Sequana) and the Marne (Matrona) in Gaul or the Boyne (Boann) in Ireland; many rivers were called simply Devona, "the Divine."

The goddess is the Celtic reflex of the primordial mother who creates life and fruitfulness through her union with the universal father-god. Welsh and Irish tradition preserve many variations on a basic triadic relationship of divine mother, father, and son. The goddess appears, for example, in Welsh as Modron (from Matrona, "Divine Mother") and Rhiannon ("Divine Queen") and in Irish as Boann and Macha. Her partner is represented by the Gaulish father-figure Sucellos, his Irish counterpart Dagda, and the Welsh Teyrnon ("Divine Lord"), and her son by the Welsh Mabon (from Maponos, "Divine Son") and Pryderi and the Irish Oenghus and Mac ind Óg, among others.

The "Divine Mother"

ZOOMORPHIC DEITIES

The rich abundance of animal imagery in Celto-Roman iconography, representing the deities in combinations of animal and human forms, finds frequent echoes in the insular literary tradition. Perhaps the most familiar instance is the deity, or deity type, known as Cernunnos, "Horned One" or "Peaked One," even though the name is attested only once, on a Paris relief. The interior relief of the Gundestrup Caldron, a 1st-century-BC vessel found in Denmark, provides a striking depiction of the antlered Cernunnos as "Lord of the Animals," seated in the yogic lotus position and accompanied by a ram-headed serpent; in this role he closely resembles the Hindu god Śiva in the guise of Paśupati, Lord of Beasts. Another prominent zoomorphic deity type is the divine bull, the Donn Cuailnge ("Brown Bull of Cooley"), which has a central role in the great Irish hero-tale *Táin Bó Cuailnge* ("The Cattle Raid of Cooley") and which recalls the Tarvos Trigaranus ("The Bull of the Three Cranes") pictured on reliefs from the cathedral at Trier, W.Ger., and at Nôtre-Dame de Paris and presumably the subject of a lost Gaulish narrative. Other animals that figure particularly prominently in association with the pantheon in Celto-Roman art as well as in insular literature are boars, dogs, bears, and horses. The horse, an instrument of Indo-European expansion, has always had a special place in the affections of the Celtic peoples. The goddess Epona, whose name, meaning "Divine Horse" or "Horse Goddess," epitomizes the religious dimension of this relationship, was a pan-Celtic deity, and her cult was adopted by the Roman cavalry and spread throughout much of Europe, even to Rome itself. She has insular analogues in the Welsh Rhiannon and in the Irish Édaín Echraidhe (*echraidhe,* "horse riding") and Macha, who outran the fastest steeds.

BELIEFS, PRACTICES, AND INSTITUTIONS

Cosmology and eschatology. Little is known about the religious beliefs of the Celts of Gaul. They believed in a life after death, for they buried food, weapons, and ornaments with the dead. The druids, the early Celtic priesthood, taught the doctrine of transmigration of souls and discussed the nature and power of the gods. The Irish believed in an otherworld, imagined sometimes as underground and sometimes as islands in the sea. The otherworld was variously called "the Land of the Living," "Delightful Plain," and "Land of the Young" and was believed to be a country where there was no sickness, old age, or death, where happiness lasted forever, and a hundred years was as one day. It was similar to the Elysium of the Greeks and may have belonged to ancient Indo-European tradition. In Celtic eschatology, as noted in Irish vision or voyage tales, a beautiful girl approaches the hero and sings to him of this happy land. He follows her, and they sail away in a boat of glass and are seen no more; or else he returns after a short time to find that all his companions are dead, for he has really been away for hundreds of years. Sometimes the hero sets out on a quest, and a magic mist descends upon him. He finds himself before a palace and enters to find a warrior and a beautiful girl who make him welcome. The warrior may be Manannán, or Lugh himself may be the one who receives him, and after strange adventures the hero returns successfully. These Irish tales, some of which date from the 8th century, are infused with the magic quality

Life after death

By courtesy of the Danish National Museum, Copenhagen

(Left) Gundestrup Caldron, from Gundestrup, Himmerland, Den., c. 1st century BC.
(Right) Interior of the caldron showing Cernunnos as "Lord of the Animals." In the Danish National Museum, Copenhagen.

that is found 400 years later in the Arthurian romances. Something of this quality is preserved, too, in the Welsh story of Branwen, daughter of Llŷr, which ends with the survivors of the great battle feasting in the presence of the severed head of Bran the Blessed, having forgotten all their suffering and sorrow. But this "delightful plain" was not accessible to all. Donn, god of the dead and ancestor of all the Irish, reigned over Tech Duinn, which was imagined as on or under Bull Island off the Beare Peninsula, and to him all men returned except the happy few.

Worship. According to Poseidonius and later classical authors Gaulish religion and culture were the concern of three professional classes—the druids, the bards, and between them an order closely associated with the druids that seems to have been best known by the Gaulish term *vates,* cognate with the Latin *vates* ("seers"). This threefold hierarchy had its reflex among the two main branches of Celts in Ireland and Wales but is best represented in early Irish tradition with its druids, *filidh* (singular *fili*), and bards; the *filidh* evidently correspond to the Gaulish *vates.*

The druids

The name druid means "knowing the oak tree" and may derive from druidic ritual, which seems in the early period to have been performed in the forest. Caesar stated that the druids avoided manual labour and paid no taxes, so that many were attracted by these privileges to join the order. They learned great numbers of verses by heart, and some studied for as long as 20 years; they thought it wrong to commit their learning to writing but used the Greek alphabet for other purposes.

As far as is known, the Celts had no temples before the Gallo-Roman period; their ceremonies took place in forest sanctuaries. In the Gallo-Roman period temples were erected, and many of them have been discovered by archaeologists in Britain as well as in Gaul.

Human sacrifice was practiced in Gaul: Cicero, Caesar, Suetonius, and Lucan all refer to it, and Pliny the Elder says that it occurred in Britain, too. It was forbidden under Tiberius and Claudius. There is some evidence that human sacrifice was known in Ireland and was forbidden by St. Patrick.

Festivals. Insular sources provide important information about Celtic religious festivals. In Ireland the year was divided into two periods of six months by the feasts

Beltine and Samhain

of Beltine (May 1) and Samhain (Samain; November 1), and each of these periods was equally divided by the feasts of Imbolc (February 1), and Lughnasadh (August 1). Samhain seems originally to have meant "summer," but by the early Irish period it had come to mark summer's end. Beltine is also called Cetśamain ("First Samhain"). Imbolc has been compared by the French scholar Joseph Vendryes to the Roman lustrations and apparently was a feast of purification for the farmers. It was sometimes called *oímelc* ("sheep milk") with reference to the lambing season. Beltine ("Fire of Bel") was the summer festival, and there is a tradition that on that day the druids drove cattle between two fires as a protection against disease. Lughnasadh was the feast of the god Lugh.

The impact of Christianity. The conversion to Christianity had inevitably a profound effect on this socio-religious system from the 5th century onward, though its character can only be extrapolated from documents of considerably later date. By the early 7th century the church had succeeded in relegating the druids to ignominious irrelevancy, while the *filidh,* masters of traditional learning, operated in easy harmony with their clerical counterparts, contriving at the same time to retain a considerable part of their pre-Christian tradition, social status, and privilege. But virtually all the vast corpus of early vernacular literature that has survived was written down in monastic scriptoria, and it is part of the task of modern scholarship to identify the relative roles of traditional continuity and ecclesiastical innovation as reflected in the written texts. Cormac's Glossary (*c.* 900) recounts that St. Patrick banished those mantic rites of the *filidh* that involved offerings to demons, and it seems probable that the church took particular pains to stamp out animal sacrifice and other rituals grossly repugnant to Christian teaching. What survived of ancient ritual practice tended to be related to *filidhecht,* the traditional repertoire of the *filidh,* or to the central institution of sacral kingship. A good example is the pervasive and persistent concept of the hierogamy (sacred marriage) of the king with the goddess of sovereignty: the sexual union, or *banais ríghi* ("wedding of kingship"), that constituted the core of the royal inauguration seems to have been purged from the ritual at an early date through ecclesiastical influence, but it remains at least implicit, and often quite explicit, for many centuries in the literary tradition. (M.D./P.Mac C.)

Germanic religion

Germanic religion comprises the mythology, religious beliefs, and cults of the Germanic-speaking peoples before their conversion to Christianity. Germanic culture extended, at various times, from the Black Sea to Greenland, or even the North American continent. Germanic religion played an important role in shaping the civilization of Europe. But since the Germanic peoples of the Continent and of England were converted to Christianity in comparatively early times, it is not surprising that less is known about the gods whom they used to worship and the forms of their religious cults than about those of Scandinavia, where Germanic religion survived until relatively late in the Middle Ages.

SOURCES

Classical and early medieval sources. The works of classical authors, written mostly in Latin and occasionally in Greek, throw some light on the religion of Germanic peoples; however, their interest in the religious practices of Germanic tribes remains limited to its direct relevance to their narrative, as when Strabo describes the gory sacrifice of Roman prisoners by the Cimbri at the end of the 2nd century BC.

For all his knowledge of the Celts, Caesar had no more than a superficial knowledge of Germans. He made some judicious observations in *Commentarii de bello Gallico* about their social and political organization, but his remarks on their religion were rather perfunctory. Contrasting Germans with the Celts of Gaul, Caesar claimed that the Germans had no druids (*i.e.,* organized priesthood), nor zeal for sacrifice, and counted as gods only the Sun, the fire god (Vulcan or *Vulcanus*), and the Moon. His limited information accounts for Caesar's assumption of the poverty of the Germanic religion and the partial inaccuracy and incompleteness of his statement.

Tacitus, on the contrary, provided a lucid picture of customs and religious practices of continental Germanic tribes in his *Germania,* written *c.* AD 98. He describes some of their rituals and occasionally names a god or goddess. While Tacitus presumably never visited Germany, his information was partly based on direct sources; he also used older works, now lost.

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

Origin of the Lombards

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

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

The first detailed document touching upon the early religion of Scandinavia is the biography by St. Rembert (or Rimbert) of St. Ansgar (or Anskar), a 9th-century missionary and now patron saint of Scandinavia, who twice visited the royal seat, Björkö, in eastern Sweden, and noticed some religious practices, among them the worship of a dead king. Ansgar was well received by the Swedes, but it was much later that they adopted Christianity.

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

German and English vernacular sources. Learned sources, such as those just mentioned, may be supplemented by a few written in vernacular in continental Germany and England. Among the most interesting are two charms, the so-called Merseburg Charms, found in a manuscript of *c.* 900, in alliterating verse. The charms appear to be of great antiquity, and the second, intended to cure sprains, contains the names of seven deities. Four of these are known from Scandinavian sources, *viz.*, Wodan (Odin), Friia (Frigg), Volla (Fulla), and Balder, but *balder* could merely designate the lord and apply to Wodan's companion Phol, an otherwise unidentified god. Sinthgunt (Sinhtgunt in the manuscript), the sister of Sunna ("Sun"), could be a name for the Moon.

A manuscript of the 9th century contains a baptismal vow in the Saxon dialect, probably dating from the 8th century. The postulant is made to renounce the Devil and all his works, as well as three gods, Thunaer (Donar/Thor), Wôden (Wodan/Odin), and Saxnôt, whose name has been associated with Seaxneat, who appears as the son of Wôden in the genealogy of the kings of Essex. Saxnôt is undoubtedly a Saxon tribal god, but it is not clear whether the second element of his name means "companion" or refers to "(sacrificial) cattle."

Old English sources

Vernacular sources in Old English are rich, but reveal little about the pre-Christian religion. The poem *Beowulf* is based upon heroic traditions, ultimately of Scandinavian origin, but in spite of its rather thorough Christianization, it retains a number of striking Germanic elements in its symbolism and contents. The fight of Beowulf against the monsters from the dark is paralleled by the struggle of Scandinavian heroes against trolls. The same heroism and defiance of death that characterize Germanic warrior ethics are found in minor historical poems, such as the *Battle of Brunanburh* and the *Battle of Maldon.* Old English literature also includes numerous charms intended as safeguards against illnesses and misfortunes, but these can hardly be called religious. In the 9th century *Runic Poem,* an old tradition about the god Ing has clearly been retained. Wôden (Odin) is also mentioned repeatedly in Old English sources; he is frequently named among ancestors of the royal houses.

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

The oldest of the sources found in the Icelandic manuscripts are in verse. Although remembered and written down in Iceland, some of these verses originated elsewhere, some in Norway and a few in Denmark and Sweden. Some of them may well be older than the settlement of Iceland, which took place toward the end of the 9th century. The Icelanders remained pagan until the year 999 or 1000.

The Icelandic manuscripts are written either in Eddic or in skaldic verse. The Eddic poetry is mostly composed in free alliterative measures, much like that of the Old English *Beowulf.* Much of it is preserved in a manuscript now called the *Elder Edda,* or *Poetic Edda,* written in Iceland *c.* 1270 and containing material centuries older. The meaning of the name *Edda* is disputed; it was not originally applied to this book but to another mentioned below.

The *Elder Edda* consists of a number of lays, which may be divided into two classes, the mythological and the heroic. The mythological poems contain stories about the northern Germanic gods; words of wisdom; a cosmogony, depicting the beginning of the world; and an apocalyptic description of the Ragnarök, the end of the ancient Scandinavian world. There is much controversy among scholars about the date and place of origin of several of the lays preserved in the *Edda* and minor collections. The first lay is the "Völuspá" ("Prophecy of the Seeress") which, in about 65 short stanzas, covers the history of the world of gods from the beginning to the Ragnarök. In spite of its clearly pagan theme, the poem reveals Christian influence in its imagery. The scenery described is that of Iceland, and it is commonly thought that it was composed in Iceland about the year 1000, when Icelanders perceived the fall of their ancient gods and the approach of Christianity.

"Hávamál"

The "Hávamál" ("Words of the High One") is a heterogeneous collection of aphorisms, homely wisdom, and counsels, as well as magic charms, ascribed to Odin. It contains at least five separate sections, some of which definitely point to their origin in Norway in the Viking age (9th–10th century) by their scenery and view of life. Of interest are the myths about Odin's erotic affairs, illustrating his cynical remarks about man's relation to woman, especially his amorous adventure leading to the theft of the precious mead. Particularly important is the account of Odin's hanging himself on the world tree, Yggdrasill, a name apparently meaning "Odin's Horse."

In another poem Odin engages in a contest of wits with an immensely wise giant (Vafthrúdnir). The poem, in the form of question and answer, tells of the cosmos, gods, giants, the beginning of the world, and its end. The other lays of the first section of the *Elder Edda* deal essentially with the adventures of the gods, especially Thor's relations with the giants, such as when he goes fetching the brewing kettle, fishing for the Midgard-Serpent, and recovering his hammer Mjölnir. The "Lokasenna" ("The Flyting of Loki"), which sharply criticizes the behaviour of the major Scandinavian gods and goddesses, perhaps on the model of Lucian's *Assembly of the Gods,* is presumably a late addition, written *c.* 1200. Similarly, the political implications in the "Rígsthula" suggest that this poem about the divine origin of social stratification dates at least to the 13th century.

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

The Icelandic and, to a lesser extent, the Norwegian manuscripts of the 13th and 14th centuries contain a great bulk of poetry of a quite different kind. This is commonly, if unjustifiably, called skaldic poetry. The skaldic verse forms were perhaps devised in Norway in the 9th century. They differ fundamentally from the traditional Germanic and Eddic forms in that the syllables are strictly counted and the lines must end in a given form. The skalds also used a complicated system of alliteration, as well as internal rhyme and consonance. With all these constraints, their short, eight-line strophes, falling neatly into four-line half strophes, are often difficult to understand because of the complexity of the syntax and of an abstruse diction, making a very extensive use of periphrastic metaphors called kennings. These phrases, *e.g.*, "Sif's hair" or "the

otter's ransom" for "gold," allude to specific myths, and their testimony is most reliable to assess pagan worship. Skaldic poetry is often composed in praise of chieftains of Norway and other Scandinavian lands. Its authors are frequently named, and their approximate date is known.

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

The *Edda* and other writings of Snorri Sturluson

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

It is likely that Snorri wrote the various sections of this book in an order opposite to that which they now have. He began with a poem exemplifying 102 different forms of verse, addressed to Haakon, the young king of Norway, and his uncle Earl Skúli Baardson. He then furnished a section entitled "Skáldskaparmál" ("Poetic Diction"), explaining and illustrating the abstruse allusions to gods and ancient heroes in the poetry of the skalds. After this, he wrote an introduction to the mythology of the north in the "Gylfaginning" ("Beguiling of Gylfi"), a section describing all of the major gods and their functions. Snorri worked partly from Eddic and skaldic poetry still extant, but partly from sources that are now lost. He presents a clear, if not altogether reliable, account of the gods, the creation of the world, and Ragnarök.

Another important work ascribed to Snorri is the *Heimskringla* ("Orb of the World"), a history of the kings of Norway from the beginning to the mid-12th century. The first section of this book, the "Ynglinga saga," is of particular interest, for in it, Snorri described the descent of the kings of Norway from the royal house of Sweden, the Ynglingar, who, in their turn, were said to descend from gods. Snorri used such written sources as were available; he also relied on skaldic poems, some of which were very old. Snorri visited Norway twice and Sweden once, and he probably used popular traditions that he heard in both countries.

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

The reliability of family sagas as sources of history has long been debated and no simple answer can be given. Each saga has to be studied separately, with a view not only to the author's sources but also to his aims. Some of the authors were antiquarians and tried to relate faithfully the history of a district, a family, or a hero; others simply entertained by writing historical fiction.

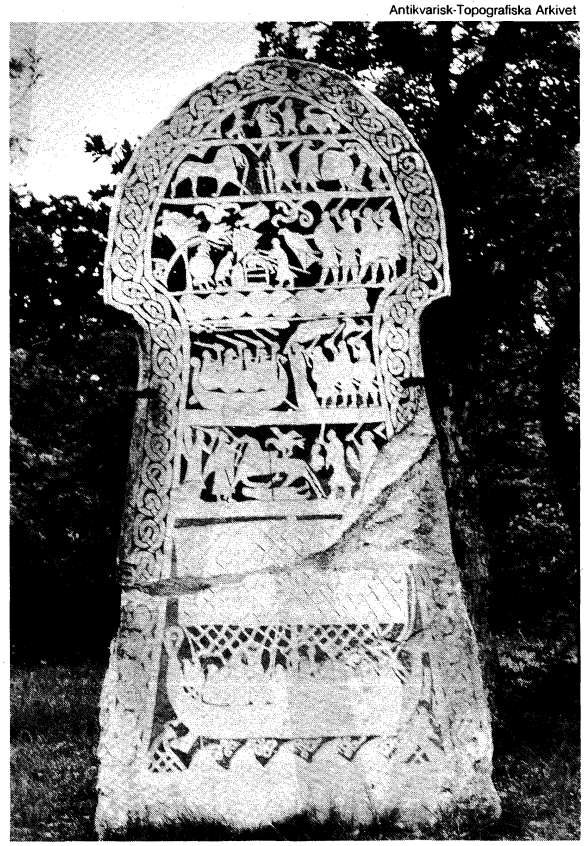
Saxo Grammaticus

About the time when the first family sagas were written, the Dane Saxo Grammaticus, secretary of Absalon, archbishop of Lund, was compiling in Latin his great history of the Danes (*Gesta Danorum*). The first nine books of this work deal with the prehistory of the Danes and are actually a history of the ancient gods and heroes. Interpreting the old religion euhemeristically (*i.e.,* by reducing the gods to the level of distinguished men), Saxo regarded the pagan gods chiefly as crafty men of old. Some of his sources may have been Danish traditions and poetry now lost, but he derived much of his information from vagrant Icelanders, of whom he speaks with some respect.

Material such as Saxo used was also used by Icelanders some generations later in the so-called heroic sagas (*Fornaldar Sögur*). Sagas of this kind describe the adventures of heroes who lived, or were supposed to have lived, in Scandinavia or on the Continent before Iceland was peopled. The gods, and particularly Odin, are frequently said to take part in the affairs of men, but since few of the heroic sagas were written before the 14th century, and the aim of their authors was often entertainment rather than instruction, these sagas can be used as sources only with utmost discrimination.

Other sources. *Archaeology.* The archaeological finds of Scandinavia are rich, and information about religious beliefs may be drawn especially from the grave goods and forms of burial. It may, in fact, be possible to trace continuity of belief from the Bronze Age to the Viking age in the 9th and 10th centuries. Archaeological finds, however, are difficult to interpret from a religious point of view. The numerous petroglyphs of southern Scandinavia, dating to the 2nd millennium BC, attest to an extensive sun cult and prevalent fertility rites. Other early Bronze Age finds such as the Trundholm chariot of the sun confirm these religious practices. Ship or boat graves were initially meant to carry the buried or cremated remains of those put in them to the otherworld, but such practices could later have become purely conventional.

Antikvarisk-Topografiska Arkivet



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

A number of small images in silver or bronze, dating from the Viking age, have also been found in various parts of Scandinavia. They show Thor with his hammer or a fertility god with full erection, perhaps Freyr; frequently found is a silver hammer, the symbol of Thor, often worn as an amulet, like the hundreds of gold medals or bracteates, representing Germanic deities worshiped on the Continent and in Scandinavia in the 5th–6th century.

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

Place-names. Theophoric place-names (derived from or

compounded with the name of a god) are found in all Germanic lands. Such names supplement the limited information available concerning pagan religion in Continental Germany and England. The theophoric place-names of Norway and Sweden are richer and have been carefully sifted. The evidence drawn from them must, however, be handled with caution. A name such as Thorslundr ("Thor's Grove") does not necessarily imply that Thor was worshiped there, for names are often transferred by settlers from one place to another, as from England to America and, in the Viking age, from the Scandinavian mainland to Iceland. Groups of theophoric place-names may, however, provide evidence of the cult of one god or another.

MYTHOLOGY

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

"Völuspá" Defective as it is, the account of the "Völuspá" appears to be the most rational description of the cosmogony. The story is told by an age-old seeress who was reared by primeval giants. In the beginning there was nothing but Ginnungagap, a void charged with magic force. Three gods, Odin and his brothers, raised up the earth, presumably from the sea into which it will ultimately sink back. The sun shone on the barren rocks and the earth was overgrown with green herbage.

Later, Odin and two other gods came upon two lifeless tree trunks, Askr and Embla, on the shore. They endowed them with breath, reason, hair, and fair countenance, thus creating the first human couple.

A quite different story is told in the didactic poem "Vafthrúdnismál" ("The Lay of Vafthrúdnir"). The poet ascribes his ancestry to a primal giant, Aurgelmir, who sometimes goes by the name Ymir. The giant grew out of the venom-cold drops spurted by the stormy rivers called Élivágar. One of the giant's legs begat a six-headed son with the other leg, and under his arms grew a maid and a youth. The earth was formed from the body of the giant Ymir who, according to Snorri, was slaughtered by Odin and his brothers. Ymir's bones were the rocks, his skull the sky, and his blood the sea. Another didactic poem, "Grímnismál" ("The Lay of Grímnir [Odin]"), adds further details. The trees were the giant's hair and his brains the clouds. Snorri quotes the three poetic sources just mentioned, giving a more coherent account and adding some details. One of the most interesting is the reference to the primeval cow Audhumla (Auðumla), formed from drops of melting rime. She was nourished by licking salty, rime-covered stones. Four rivers of milk flowed from her udders and thus she fed the giant Ymir. The cow licked the stones into the shape of a man; this was Buri (Búri), who was to be grandfather of Odin and his brothers. The theme of the creation of the world from parts of the body of a primeval being is also found in Indo-Iranian tradition and may belong to the Indo-European heritage in Germanic religion.

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

The gods. Old Norse sources name a great number of deities. The evidence of place-names suggests that one cult succeeded another. Names, especially those in southeastern Norway and southern Sweden, suggest that there was once widespread worship of a god Ull (Ullr). Indeed, an early poem reports an oath on the ring of Ull, suggesting that he was once one of the highest gods, at least in some areas. Beyond that, little is known about Ull; he was god of the bow and snowshoes, and, according to Saxo Grammaticus, who calls him Ollerus, he temporarily replaced Odin when the latter was banned from his throne.

The gods can be divided roughly into two tribes, Aesir and Vanir. At one time, according to fairly reliable sources, there was war between the Aesir and the Vanir, but when neither side could score a decisive victory they made peace and exchanged hostages. In this way, the specialized fertility gods, the Vanir, Njörd (Njörðr), his son Freyr, and presumably his daughter, Freyja, came to dwell among the Aesir and to be accepted in their hierarchy. Two tribes of gods: Aesir and Vanir

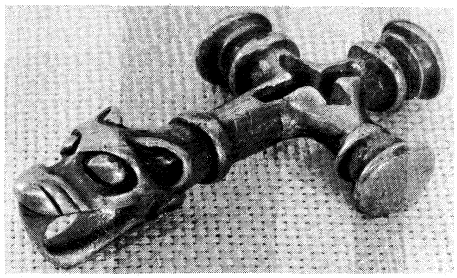
Odin (*Óðinn*). According to literary sources, Odin was the foremost of the Aesir, but the limited occurrence of his name in place-names seems to indicate that his worship was not widespread. He appears, however, to have been the god of kings and nobility more than the deity to whom the common man would turn for support. His name defines him as the god of inspired mental activity and strong emotional stress, as it is related to Icelandic *óðr,* which applies to the movements of the mind, and to German *Wut,* meaning "rage," or "fury." This qualifies him as the god of poetic inspiration and the stories about the origin of poetry narrate how Odin brought the sacred mead of poetry to the world of the gods. This beverage was first brewed from the blood of a wise god, Kvasir, who was murdered by dwarfs. It later came into the hands of a giant and was stolen by Odin, who flew from the giant's stronghold in the shape of an eagle, carrying the sacred mead in his crop to regurgitate it in the dwelling of the gods. Therefore, the early skalds designate poetry as "Kvasir's blood" or "Odin's theft."

There is also a darker side to Odin's personality: he incites kinsmen to fight and turns against his own favourites, because he needs heroes in the otherworld to join him in the final battle against the forces of destruction at the time of Ragnarök. Therefore, the fallen warriors on the battlefield are said to go to his castle Valhalla (Valhöll), the "Hall of the Slain," where they live in bliss, training for the ultimate combat. He is also a necromancer and a powerful magician who can make hanged men talk. He is the god of the hanged, because he hanged himself on the cosmic tree Yggdrasill to acquire his occult wisdom. As the "Hávamál" tells us, he hung there for nine nights, pierced with a spear, sacrificed to himself, nearly dead, to gain the mastery of the runes and the knowledge of the magic spells that blunt a foe's weapons or free a friend from fetters. Valhalla

Odin could change his shape at will, and, with his body in cataleptic sleep, he traveled to other worlds, like a shaman. As god of the dead, he was accompanied by carrion beasts, two wolves and two ravens. These birds kept him informed of what happened in the world, adding to the knowledge he had acquired by relinquishing his one eye in the well of Mímir under the tree Yggdrasill.

Untrustworthy, Odin may break the most sacred oath on the holy ring. As "spear-thruster," he opens the hostilities, and in the bellicose period of the Viking expeditions his cult appeared to gain momentum. Odin, like Wôden or Wotan, is, however, essentially the sovereign god, whom the Germanic dynasties, in England as well as in Scandinavia, originally regarded as their divine founder. He thus maintains the prominent position of Wōðan[az] in classical antiquity, to whom, according to Tacitus, human sacrifice was offered. Latin writers identified Wōðan[az] with Mercury, as the name of the day, Wednesday, (*i.e.,* "day of Wôden"), for *Mercurii dies* (French *mercredi*), indicates. It is possible that the tribal god of the Semnones, described by Tacitus as *regnator omnium deus* ("the god governing all"), could be identified with Wōðan[az]. They would indeed sacrifice a man to him in a sacred grove in what the ancient author describes as a "horrendous ritual."

Thor (*þórr*). Thor is a god of very different stamp. Place-names, personal names, poetry, and prose show that he was worshiped widely, especially toward the end of the pagan period. Thor is described as Odin's son, but his name derives from the Germanic term for "thunder." Like Indra and other Indo-European thunder-gods, he is essentially the champion of the gods, being constantly involved in struggles with the giants. His main weapon is a short-handled hammer, Mjölnir, with which he smashes the skull of his antagonists. One of his best-known adventures describes his pulling the cosmic serpent Jörmungand (Jörmungandr), which surrounds the world, out of the



Silver image showing the integration of pagan (Thor's hammer) and Christian (cross) symbols; found in southern Iceland. In the National Museum, Reykjavík.
By courtesy of the National Museum, Reykjavík; photograph, Gisli Gestsson

ocean. As he fails to kill the monster then, he will have to face it again in a combat to the finish in which they both die, in the Ragnarök.

Thor is the god of the common man. As place-names in eastern Scandinavia and in England indicate, peasants worshiped him because he brought the rains that ensured good crops. Warriors trusted him, and he seems to have been popular with them everywhere. He was well known as Thunor in the Saxon and Jutish areas in England; the Saxons on the mainland venerated him as Thunær. When the Vikings conquered Normandy and the Varangians settled in Russia, they called upon Thor to help them in their military enterprises.

On account of his association with thunder, the Germanic god *þunraz* (Thor) was equated with Jupiter by the Romans; hence, the name of the day, Thursday (German *Donnerstag*), for *Jovis dies* (Italian *giovedì*). Thor traveled in a chariot drawn by goats, and later evidence suggested that thunder was thought of as the sound of his chariot.

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

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

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

There has been much dispute among scholars about the symbolic significance of Balder's myth. He has been described as a dying spring god; some have stressed his Christ-like features in the west Norse version. The major protagonists in the drama have warrior names, and the game in which the gods hurl missiles at the almost invulnerable Balder is reminiscent of an initiatory test.

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

Loki deceived the gods and cheated them, but sometimes he got them out of trouble. He is seen in company with Odin and an obscure god Hœnir, and he is called the friend of Thor. He is essentially a "trickster" figure who can change sex and shape at will. Thus, he can give birth as well as beget offspring. The eight-legged horse of Odin, Sleipnir, was born of Loki in the shape of a mare. According to an Eddic lay, Loki ate the heart of an evil woman and grew pregnant. He fights with Heimdall in the shape of a seal for the possession of the Brísingamen necklace, and later, he sneaks into Freyja's residence in the form of a fly to steal the same precious object for Odin. According to an early poem, Odin and Loki had mixed their blood as foster brothers. It has been suggested that Loki was a hypostasis of Odin, or at least that he represents Odin's darkest side. He seems to symbolize "impulsive intelligence," together with an irrepressible urge to act and an unpredictable maliciousness.

Origin of monsters

Minor Aesir. A number of minor deities are also ranked among the Aesir. The god Heimdall (Heimdal[l]r) is particularly interesting, but rather enigmatic. His antagonism with Loki, with whom he struggles for the possession of the Brísingamen necklace, results in their killing each other in the Ragnarök, according to Snorri. Heimdall is of mysterious origin: he is the son of nine mothers, said to be sisters, all of whom bear names of giantesses, though they are mostly identified with the storm waves. Heimdall lives in Himinbjörg ("Heavenly Fells"), at the edge of the world of the Aesir, which he guards against the giants. He is endowed with a wonderful hearing, detecting anything in the world, but he is blamed with drinking too much mead. When the Ragnarök draws near, he will blow his ringing horn (Gjallarhorn).

Another myth in which he appears as Rigr (Rígr), a name probably derived from the Irish *rí* ("king"), makes Heimdall the father of mankind. He consorted with three women, from whom descend the three classes of men—serf (*thrall*), freeman (*karl*), and nobleman (*jarl*).

Information about the Scandinavian gods is based chiefly on poetry composed late in the pagan period and on the remarks of outside observers, who generally had little interest in what they considered to be heathendom. Many gods were nearly forgotten when these authors mentioned them, as is the case with Ull, described above. Similarly, memories had apparently faded about Tyr (Týr), who must have been a major god in early times. His name, derived from Germanic *Tîwaz* (Old English *Tīw*) and related to the Greek god Zeus, suggests that he was originally a sky-god, but in Roman times, he was equated with Mars, and hence *dies Martis* (Mars's day; French *mardi*) became Tuesday (Icelandic *Týs dagr*). Tyr is the one-handed god, because one of his hands had been bitten off by the wolf Fenrir. He is brave and warlike; in the Ragnarök he will face the hellhound Garm (Garmr), and they will kill each other. Like other gods, Tyr is said to be a son of Odin, but, according to one early poem, he was the son of a giant. Tyr's cult is remembered in place-names, particularly those of Denmark.

Bragi. Bragi appears in later sources as the god of poetry and eloquence. It is remarkable that the first recorded skald, living in the 9th century, was also called Bragi. Since there is no record of a cult of the god Bragi, some have suspected that the god and the poet are identical.

Frigg. Frigg is the wife of Odin. In the southern Germanic sources she appears as Friia (Second Merseburg Charm) or Frea (Langobardic), the spouse of Wodan. Snorri depicted her as the weeping mother of Balder, but Saxo described her as unchaste and makes her misconduct responsible for the temporary banishment of Odin. In the "Ynglinga saga," Odin's brothers Vili and Vé share her during his absence in a polyandric relationship similar to that of Draupadī in Hindu myth. She has been equated with Venus, and her name survives in Friday (Old English *Frīgedæg*) from *dies Veneris,* Venus' day.

Idun (*Iðunn*). According to an early skaldic poem (*c.*

900), Idun, the wife of Bragi, was entrusted with the apples that prevent the gods from growing old. She was abducted by the giant Thjazi, but Loki brought her back with the precious apples. This myth has many parallels such as Heracles' obtaining the golden apples of the Hesperides.

Jörd (Jörðr). The name Jörd means "earth," but this goddess who is described as the mother of Thor, and consequently Odin's lover, is also known under different names, such as Fjörgyn ("Earth"), perhaps originally a goddess of the furrow, and Hlódyn (Hlóðyn). A *dea Hludana* is also remembered in votive inscriptions of lower Germany and Holland.

The Vanir. The Vanir represent a distinct group of gods associated with wealth, health, and fertility. Although they would also fight, the Vanir were not essentially gods of battle, like the Aesir. The best known Vanir—Njörd, Freyr, and probably Freyja—came as hostages to the Aesir. Njörd was the father of the god Freyr and the goddess Freyja.

In his *Germania,* Tacitus described the worship of a goddess, Nerthus, on an island, probably in the Baltic Sea. Whatever symbol represented her was kept hidden in a grove and taken around once a year in a covered chariot. During her pageant, there was rejoicing and peace, and all weapons were laid aside. Afterward, she was bathed in a lake and returned to her grove, but those who participated in her lustration were drowned in the lake as a sacrifice to thank her for her blessings.

Njörd

Nerthus is described as Terra Mater ("Earth Mother"), but her name corresponds to that of the god Njörd (from Germanic *Nerthuz*). Scholars have attempted various explanations of this puzzling change of sex, assuming that the original deity was androgynous or claiming that the loss of feminine nouns of the type *Njörd* represents triggered the reinterpretation of the goddess as a male god. As Njörd is essentially a god of the sea and its riches, it may be preferable to consider Nerthus and Njörd as originally separate gods altogether, whose relationship might be similar to that of Poseidon ("Husband of the Earth-Goddess") and Demeter ("Earth Mother") in Arcadia. Etymologically, the name *Njörd* could then be related to that of the Greek "Old Man of the Sea," Nereus. Before coming to the Aesir, Njörd was supposed to have begotten his two children with his (unnamed) sister. Since such incestuous unions were not allowed among the Aesir, Njörd afterward married Skadi (Skaði), daughter of the giant Thjazi. Evidence from place-names shows that Njörd was worshiped widely in Sweden and Norway, and he was one of the gods whom Icelanders invoked when they swore their most sacred oaths.

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

The Eddic poem *Skírnismál* ("The Lay of Skírnir") relates the wooing of Freyr's bride, Gerd (Gerðr), a giant-maiden. This story has often been considered as a fertility myth. Gerdr (from *garðr,* "field") is held fast in the clutches of the frost-giants of winter. Thus, Freyr, as sun-god, would free her. However, this interpretation rests entirely on disputable etymologies. The narrative indicates that Freyr's bride belongs to the otherworld, and her wooing may rather symbolize the affinities of the fertility god with the chthonian powers, dominating the cycle of life and death. Several animals were sacred to Freyr, particularly the horse and, because of his great fertility, the boar.

The centre of Freyr's cult was Uppsala, and he was once said to be king of the Swedes. His reign was one of peace and plenty. While Freyr reigned in Sweden, a certain Frodi ruled the Danes, and the Danes attributed this age of prosperity to him. Frodi (Fróði) was also conveyed ceremoniously in a chariot, and some have seen him as no other than a doublet of Freyr. Freyr was said to be ancestor of the Ynglingar, the Swedish royal family. Such myths are connected with the concept of "divine kingship" in the Germanic world, but earlier views on "sacral royalty" are now being challenged.

Freyja. Freyr's sister, Freyja, shares several features with her brother. She was the goddess of love, wealth, and fertility. She owned precious jewels such as the famous Brísingamen necklace, forged by dwarfs. She is said to be weeping tears of gold for her absent husband, but she is also blamed for being promiscuous. She practiced a disreputable kind of magic, called *seiðr,* which she taught Odin. She was known under various names, some obscure such as Mardöll, and others, such as Sýr ("Sow"), referring to her association with animals. Taking half of those who fall in battle, Freyja had some affinity with the chthonian deities of death.

This relation of fertility goddesses with the otherworld is already illustrated by the Germanic mother goddesses or *matronae,* whose cult was widespread along the lower Rhine in Roman imperial times. They are often represented with chthonian symbols such as the dog, the snake, or baskets of fruit. The same applies to the goddess Nehalennia, worshiped near the mouth of the Scheldt River. Her name may be related to Greek *nekués,* "spirits of the dead."

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

Dísir and *fylgjur*

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

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

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

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

Dwarfs. Dwarfs (*dvergar*) play a part in Norse mythology. They were very wise and expert craftsmen who forged practically all of the treasures of the gods, in particular Thor's hammer. Snorri said that they originated as maggots in the flesh of the slaughtered giant Ymir. Four of them are supporting the sky, made of the skull of this primeval giant. They may have been originally nature spirits or demonic beings, living in mountain caves, but they generally were friendly to man.

BELIEFS, PRACTICES, AND INSTITUTIONS

Worship. Sacrifice often was conducted in the open or in groves and forests. The human sacrifice to the tribal god of the Semnones, described by Tacitus, took place in a sacred grove; other examples of sacred groves include the one in which Nerthus usually resides. Tacitus does,

however, mention temples in Germany, though they were probably few. Old English laws mention fenced places around a stone, tree, or other object of worship. In Scandinavia, men brought sacrifice to groves and waterfalls.

A common word for a holy place in Old English is *hearg* and in Old High German *harug,* occasionally glossed as *lucus* ("grove") or *nemus* ("forest"). The corresponding Old Norse word, *hörgr,* denotes a cairn, a pile of stones used as an altar; the word was also used occasionally for roofed temples. Another term applied to sacred places in Scandinavia was *vé* (compare with *vígja,* "to consecrate"), which appears in many place-names; *e.g.,* Odense (older Óðinsvé).

Temples

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

The word *hof,* commonly applied to temples in the literature of Iceland, seems to belong to the later rather than to the earlier period. A detailed description of a *hof* is given in one of the sagas. The temple consisted of two compartments, perhaps analogous to the chancel and the nave of a church. The images of the gods were kept in the chancel. This does not imply, however, that Icelandic temples of the 10th century were modeled on churches; rather they resembled large Icelandic farmhouses. A building believed to be a temple has been excavated in northern Iceland, and its outline agrees closely with that described in the saga.

Temples on the mainland of Scandinavia were probably built of wood, of which nothing survives, although an influence of pagan temples may be discernible in the so-called stave churches. At the close of the pagan period, the most splendid temple of all was at Uppsala. It was richly described by Adam of Bremen, whose report is based on statements of eyewitnesses, though he may have been influenced by the biblical description of Solomon's temple. Statues of Thor, Wodan, and Fricco (Freyr) stood together within it; the whole building was covered with gold, which could be seen glittering from afar. There were also famous temples in Norway, but no detailed descriptions are given of them.

Sacrifice took different forms. Roman authors repeatedly mention the sacrifice of prisoners of war to the gods of victory. The *thralls* who bathed the numen of Nerthus paid for the revelation of her secret identity with their lives. A detailed description of a sacrificial feast is given in a saga about a king of Norway. All kinds of cattle were slaughtered, and blood was sprinkled inside and out; the meat was consumed and toasts were drunk to Odin, Njörd, and Freyr. The most detailed description of a sacrifice is that given by Adam of Bremen. Every nine years a great festival was held at Uppsala, and sacrifice was conducted in a sacred grove that stood beside the temple. The victims, human and animal, were hung on trees. One of the trees in this grove was holier than all the others and beneath it lay a well into which a living man would be plunged.

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

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

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

The presence of ships or boats in graves, and occasionally of chariots and horses, may suggest that the dead were thought to go on a journey to the otherworld, but this is questionable; such accoutrements more likely reflected a person's earthly occupation. Some records imply that the dead needed company; a wife, mistress, or servant would be placed in the grave with them. The famous Oseberg grave contained the bones of two women, probably a queen and her servant. Some stories suggest the existence of an ancient belief in rebirth, but a medieval writer labels the notion an old wives' tale. On the whole, beliefs in afterlife seem rather gloomy. The dead pass, perhaps by slow stages, to a dark, misty world called Niflheim (Niflheimr).

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

Cosmic destruction: Twilight of the Gods

Only a brief summary of this rich subject can be attempted here. Through their own work, and especially because of the strength of Thor, gods have kept the demons of destruction at bay. The savage wolf Fenrir is chained, as is the guileful Loki, but they will break loose. Giants and other monsters will attack the world of gods and humans from various directions. Odin will fight the wolf and lose his life, to be avenged by his son Vidar (Víðarr), who will pierce the beast to the heart. Thor will face the World Serpent, and they will kill each other. The sun will turn black, the stars vanish, and fire will play against the firmament. The earth will sink into the sea but will rise again, purified and renewed. Unsown fields will bear wheat. Balder and his innocent slayer, Höd, will return to inhabit the dwellings of gods. Worthy people will live forever in a shining hall thatched with gold.

Although the cosmic cataclysm portrayed by the poet of the "Völuspá" reflects the apocalyptic imagery of the Book of Revelation, it is essentially a symbolic reflection of the waning Germanic world, ineluctably moving to its destruction because of the outrages committed by its divine and human representatives. According to another Eddic poem, the wolf will swallow Odin and, in revenge, his son will tear the jaws of the beast asunder. Several more details are given in other sources, generally cruder than those of the "Völuspá."

THE END OF PAGANISM

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

The story of the conversion of Iceland is known best because of the wealth of historical documents written in that country during the Middle Ages. Icelanders were, in many ways, the most international of northern Scandinavians. Among those who settled in Iceland in the late 9th century were men and women partly of Norse stock from Christian Ireland. Some of these were Christians; some were mixed in their beliefs, worshiping Christ and Thor at once. There were others who believed in no gods at all. Lack of faith in the heathen gods seems to have grown during the 10th century. Influence of Christian thought on some Icelandic poets is noticeable. Occasional missions to Iceland in the later 10th century are recorded, but little progress was made until Olaf I Tryggvason, king of Norway, sent out the German priest Thangbrand *c.* 997. Thangbrand was a ruthless, brutal man; he was outlawed and returned to Norway *c.* 999. But in the year after Thangbrand left (*c.* 1000), the Icelandic parliament (Althingi) resolved, at the instigation of King Olaf, that all should be baptized, although concessions were made to those who wished to practice heathen rites in private. Many of those who had been hereditary pagan chieftains became leaders of the church and, largely for this reason, tradition survived in Iceland as in no other Scandinavian land.

Conversions to Christianity

The conversion of Norway was far less peaceful. Much is known about it, chiefly from highly colourful Icelandic records. Olaf Tryggvason, who had come to Norway from England *c.* 995, quickly overcame the arch-pagan ruler Haakon Sigurdsson. Paganism was deeply rooted in the

minds of hereditary landowners, as the whole social system was largely founded upon its principles. Using fire and sword rather than persuasion, Olaf converted the whole of Norway in his short reign of five years. When he died in a naval battle, *c.* 1000, many of Olaf's subjects were Christians in name only. By the time Olaf II Haraldsson (later Saint Olaf) came to the throne about 15 years later, some of the Norwegians were baptized and some not, and everyone believed whatever he chose. Olaf II set out to complete the work of his predecessor, resorting to the same methods. He was such a tyrant that his own subjects, Christian though they were, drove him into exile in Russia. When he returned with a motley army, *c.* 1030, he met his death and was soon regarded as a saint. For all his faults, Olaf had established Christianity firmly in Norway.

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

The picture that Scandinavian sources provide of Germanic religion is to a large extent lopsided, since many of the documents date to the period when waning paganism was threatened with doom by the growing impact of Christianity. This may account for the pessimistic worldview that pervades some aspects of Eddic poetry, as well as for some rather derogatory descriptions of the behaviour of the gods. The rigorous ethics of early Germanic society, based on trust, loyalty, and courage, and the perhaps somewhat idealized picture of the moral code given by Tacitus, had a divine sanction, but when Christianity arrived in the north, the message had apparently been dimmed by the gods' disrespect of their most solemn oaths. Paganism no longer had the stamina and inner drive to resist the pressure of Christianity, with its strong, well-organized church and its positive monotheistic creed, encompassing faith and ethics. (E.O.G.T.-P./E.C.Po.)

Finno-Ugric religion

The religion of the Finno-Ugric peoples, who inhabit regions of northern Scandinavia, Siberia, the Baltic area, and central Europe, is an admixture of agrarian and nomadic primitive religion and of Christianity and Islām. This treatment is concerned primarily with the pre-Christian and pre-Islāmic elements of Finno-Ugric religion.

GEOGRAPHIC AND CULTURAL BACKGROUND

The Finno-Ugric peoples. The area inhabited by the Finno-Ugric peoples is extensive: from Norway to the region of the Ob River in Siberia and southward into the Carpathian Basin in central Europe and the Ukraine. The history of their geographic dispersion is based almost entirely on linguistic criteria, since historical knowledge is recent and archaeological finds are scanty and interpreted variously.

Geographic distribution

The Finno-Ugric languages and the Samoyed languages together form the Uralic family of languages, which began to split up about 4000–5000 BC. The original Uralic people are thought to have lived in the region between the Ural Mountains and the middle reaches of the Volga River. Their descendants in the north are the Nenets, who live on the shores of the Arctic Ocean between the Taymyr and the Kanin peninsulas. In the south, the original speakers of the parent Finno-Ugric language probably began to disperse by 3000 BC, when the Ugrians formed their own group. One branch moved northeast, behind the Ural Mountains: the Ostyak (who in their own language call themselves Khanty), living east of the Ob River, and the Vogul (who call themselves Mansi), living west of the Ob River. The other branch spread southward and made contact with the Bulgar Turks and the Khazars; in the year 895 this branch (the Magyars [Hungarians]), together with certain Turkish tribes, conquered what is now Hungary. In this way, the largest, but at the same time linguistically the most isolated, Finno-Ugric nation came into existence. Other Magyars live in the countries of Romania and Czechoslovakia.

The Permian branch of the Finno-Ugric populations living in central Russia split from the other groups between 2500 and 2000 BC; the linguistic differentiation is not very great between the present-day Permians, who are divided into Votyaks (called Udmurts, living between the Kama and Vyatka rivers) and Zyryan (called Komi, living in the region between the upper reaches of the Western Dvina River, Kama, and Pechora)—the differentiation only occurred a little over 1,000 years ago. An intermediary group between the two branches are the Permyaks, whose language is sometimes considered a dialect of Zyryan.

Farther to the south, the differentiation of the Volga Finns into separate groups probably began about 1200 BC. The Volga Finns consist today of the Mordvins (including the Moksha in the southeast and the Erzya in the northwest), living in a rather large region near the middle reaches of the Volga River, and the Cheremis (the Mari), living in the vicinity of the confluence of the Volga and the Kama.

When the Baltic Finns came to the regions bordering on the Baltic Sea is not certain. The latest possible date would be *c.* 1500 BC (the evidence being the Baltic loan words in proto-Finnic), when the "proto-Finns" still maintained contact with the Mordvins and the Lapps. A much earlier date is possible, however, as there must have been many and repeated migrations by the Finno-Ugric populations westward from the Ural Mountains toward the Baltic regions. Initially, settlement was sparse, as is always the case with hunting cultures, but language differentiation sped up with the change to sedentary agriculture. The Lapps (called the Sāpmi) have been the slowest of the Finno-Ugric peoples to relinquish the hunting and nomadic culture—which has withdrawn slowly toward the north—and they themselves have moved from the direction of Lakes Ladoga and Onega (northeast of Leningrad) to the northern parts of Fennoscandia and the Kola Peninsula (northwestern Soviet Union).

After separating from early proto-Finnic about 3,000 years ago, the Lapp language became divided into a number of very different dialects. The oldest population settlements of the Baltic Finns were to the south of the Gulf of Finland and to the south of Lake Ladoga. The most westerly group, the Livonians (in the north of Courland, now part of the Latvian S.S.R.), is disappearing. The Estonians (living in the Estonian S.S.R.) are one of the three most advanced of the Finno-Ugric peoples, the others being the Finns and the Hungarians. Small but interesting cultures are represented by the Greek Orthodox Votes and Izhora Ingrians, both nearly extinct groups living near the head of the Gulf of Finland in an area once called Ingria, the Veps (living near Lake Onega), and the Karelians (living in central Russian S.F.S.R., the Karelian A.S.S.R., and Finland), as well as the Ludes in Olonets, who speak a transition dialect. The population moved into Finland from the south and southeast.

Ecological and intercultural factors. To attain a proper understanding of the history and phenomenology of the religion of the Finno-Ugric peoples, two basic influences must be borne in mind: the ecological factors and the pressure of alien cultures on the original religious tradition. The result of both factors has been a great variation in the religious atmosphere in different places.

Influences on Finno-Ugric religion

The Lapps, Nenets, Vogul, and Ostyak—who all have been associated with a nomadic and hunting culture in Arctic regions—retain a religious life that has many ancient elements. The Finns, Karelians, and Zyryans have practiced hunting up to the present, but they have been familiar with agriculture for thousands of years. The peoples on the south side of the Gulf of Finland, such as the Estonians, have long practiced agriculture and cattle breeding as well as fishing, but hunting has not been as important to them. The Finno-Ugric peoples of the southeast, like the Votyaks and the Cheremis, have practiced agriculture and cattle breeding only. The agrarian economy of the Hungarians, with its seminomadic features, is the outcome of a complicated history.

Habitat, climate, and other ecological factors have had

an important influence on economy and social organization and on traditional religion. Some of the differences between the various Finno-Ugric peoples, however, can be traced to outside cultural influences. The southeastern Finno-Ugric peoples have been marked by Turko-Tatar influence. In the 8th century the Votyaks and the Cheremis came under Bulgar domination; the conversion of the Bulgars to Islām in 922 and the subsequent Tatar domination in eastern Russia (1236–1552) gave added significance to the Arab–Islāmic tradition. In the 16th and 17th centuries, the Volga Finns, the Permians, the Ob Ugrians, and the Nenets finally came under the domination of Moscow; before this, Orthodox missionaries had worked, for example, among the Zyryans (St. Stephen, 14th century) and the Baltic Finns.

The influence of Slavic tradition on the Finno-Ugric peoples has been considerable—from the point of view of both folk religion and the more institutionalized Orthodox faith, though some of this influence in many places is late and superficial. There are also Finno-Ugric substrates in the Russian tradition in the north and northwest of Russia. Pre-Christian practices were still alive in the early 20th century, and among the Votyaks, the Ob Ugrians, and the Nenets there were still people who were unbaptized. Roman (Catholic) and Byzantine (Orthodox) traditions met one another in Finland and Estonia, but the Orthodox groups remain established only in the eastern regions. Most of Finland was converted to Christianity by way of Sweden, beginning in the 12th century, and the country remained Roman Catholic until Lutheranism was established in the 16th century. The position of the Hungarians, who formed a pocket surrounded by alien cultures, resulted in an extremely mixed array of contacts at different levels.

Thus, each of the Finno-Ugric peoples has its own cultural history, habitat, and level of civilization. In considering their religion, all this must be borne in mind. The Hungarians, Finns, and Estonians have the longest literary traditions, while a number of the other peoples are only now developing written literature in their own language. Ancient popular belief, preserved in oral tradition, has for the most part developed more persistently on the periphery, but near centres of culture it has become a minor growth alongside institutional religions.

Difficulties involved in the study of Finno-Ugric religion

The problem of the concept of a Finno-Ugric religion. Since it is not possible to find a single formula to cover Finno-Ugric cultures and religions and since the relationship between the peoples is often distant both geographically and historically, it may well be asked whether there is any utility in attempting, by means of comparative methods, to discover some common or basic substratum in Finno-Ugric religion. Many earlier scholars attempted this enthusiastically, but today there is general agreement that a hypothetical reconstruction representing the "original religion" of a single language family is virtually impossible. That ancient tradition may have been preserved in different regions, although fragmented and adapted to new conditions, is, of course, possible, and indeed seemingly trustworthy discoveries have been made that substantiate this view. One must, however, be extremely circumspect in projecting hypotheses applying to the entire linguistic group. Genetic-historical considerations are of great importance when dealing with those areas of the language family where a cultural connection has subsisted long and late.

The search for a common historical tradition is not, however, the most rewarding aspect of the study of Finno-Ugric religions. The religio-phenomenological approach is equally interesting and significant. In the course of conducting nonhistorical studies of similarities and differences in Finno-Ugric religious material, scholars have uncovered a spectrum of basic religious forms running from Arctic hunting and fishing cultures to southern cattle breeding and agriculture.

MYTHOLOGY

Creation, cosmography, and cosmology. The most widespread account of the creation among the Finno-Ugric peoples is the earth-diver myth. In the north it is known in an area extending from eastern Finland to the Ob River, and in the south it is found, for example, among the Mordvins. This myth, which is well known in North America and Siberia, is fairly constant in form among the Finno-Ugric peoples. In the Mordvin variant, God sits on a rock in the middle of the primeval sea and spits into the water; the saliva begins to grow and God strikes it with a staff, whereupon the Devil comes out of it (sometimes in the form of a goose). God orders the Devil to dive into the sea for earth from the bottom; at the third attempt, he succeeds but tries to hide some of the earth in his mouth. While God scatters sand, the earth begins to grow and the Devil's deceit is unmasked, and the earth found in his cheek becomes mountains and hills. The eastern Finnish myth contains an interesting detail; God stands on the top of a golden statue and orders his reflection on the water to rise, and this becomes the Devil.

Etiological (explanatory and expanding) continuations of the basic myth are common; the Devil demands for himself a piece of earth the size of the end of a stick, and from the hole that results vermin emerge—mice, fleas, mosquitoes, flies, and other such living things. Indo-Iranian influence has been seen in the dualism of the myth—setting God against the Devil—since religious dualism is most significant in Indo-Iranian religion. A water bird may be older than the Devil; it also occurs, however, without the dualistic emphasis. Thus, in an account by the Yenisey Ostyak, the great shaman (a medicine man with psychic abilities) Doh glides above the primeval sea among the water birds, asks the red-throated loon to dive for earth from the bottom of the sea, and with the earth makes an island. A rarer, but apparently ancient, myth is found among the Vogul: the god of the skies lets earth come down from heaven and places it on the surface of the great primeval sea.

Cosmic egg myth

The world made from an egg is a myth best known in equatorial regions, though the most northerly points of its distribution are in Finland and Estonia. A water bird or an eagle makes its nest on the knee of the creator (Väinämöinen), who is floating in the water; it lays an egg, which rolls into the water, and pieces of it become the earth, the sky, the moon, and the stars. Myths concerning the creation of man are found in the north among the Vogul and in the south among the Volga Finns. The common element among all such myths is that man, on the brink of achieving perfection, had his hairy covering transferred to the dog by the Devil, whose spit blighted man and made him subject to disease and death. In Finland the variant of yet another anthropogonic (origin of man) myth has been found: a hummock rises from the sea, a tree stump thereon splits open, and the first human couple steps forth.

Finno-Ugric cosmographic (world-describing) concepts include the following well-known mythological themes: a stream or sea encircling the round world; a canopy of the heavens, the central point of which is the North Star (a kind of nail on which the sky rotates); a world pole supporting the sky; a world mountain and a world tree rising in the middle of the earth; animals carrying the earth; and the nub of the earth and the nub of the sea (an abyss that swallows ships). From these and from other materials more or less coherent cosmographies have been formed in different places; the central components are the sky, the earth, and the underworld. Among the Ob Ugrians and the Nenets is found a myth of the seven- or nine-storied heaven.

The cosmogonic (concerning the origin of the world) and cosmographic myths have had important ritual functions and have provided the basis for cosmology (the ordering of the world). When, in incantations and prayers, numerous natural, cultural, and social phenomena derive from these basic myths, it is not a matter of giving an explanation but of finding the connection with the decisive primeval events that gave the world its lasting order. A pillar representing the world pole has been worshiped by the Lapps and the Ob Ugrians, especially as a symbol of the world order.

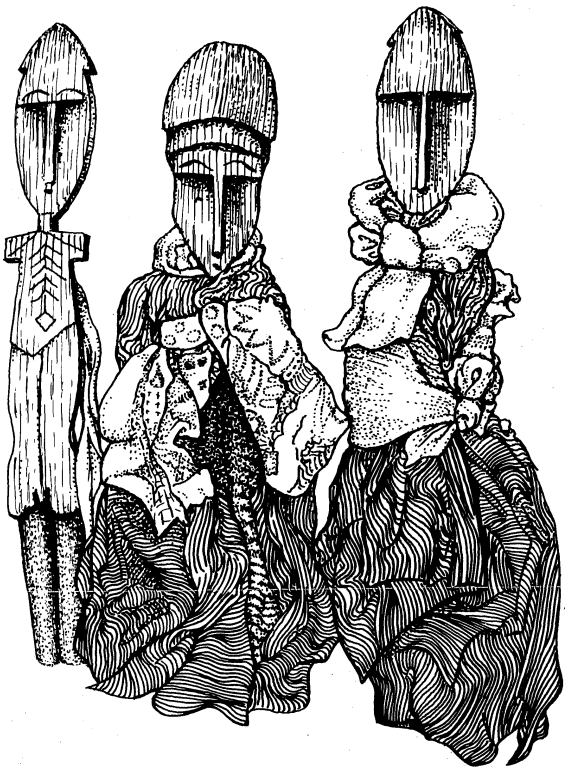
High gods. The semantic elements "sky" and "god of the sky" are found to be so close in the terminology of certain of the Finno-Ugric peoples (for example,

Cheremis Jumo, Finnish Jumala, Votyak Inmar, Zyryan Jen, Nenets Num) that the association cannot be a recent phenomenon. The tradition of the god of the sky is many-layered, and the influence of monotheism, especially of Christianity and Islām, is widely exhibited. This influence was evidently preceded by that of ancient southern high cultures. Thus the Cheremis Jumo has a real court with servants in his heaven, and these servants act as intermediaries between humans and the god of the sky. This indicates a Turko-Tatar influence, which can also be seen in the Votyak Inmar; Christian elements, however, are also found (Inmar's mother is related to the Virgin Mary). "Great," the most common epithet for Inmar and Jumo, reminds one of Allāh. The Mordvin god of the sky (Škaj, "creator" or "birth giver," among the Moksha people, and also Ńiškė-pas, "the great inseminating god") is the chief of the gods, all-knowing and all-seeing, who is not approached for trivial things. He appears, however, very concretely in a festival connected with the beginning of the spring plowing. In this festival an old man represents the god of the sky and from an attic or a tree answers questions put to him by people who pray about health, the grain harvest, the weather, and other matters. The gods of the sky of the Arctic Finno-Ugric peoples (Nenets Num; Ostyak Num-Turom and Sängke; Vogul Num-Tarom; Lappish Tiermes, Horagalles, and Radien) are the high gods of hunting and nomadic cultures, who sometimes appear in myths as creator gods and culture heroes (often as *dei otiosi,* or "inactive gods," without a cult) and sometimes as venerated gods of the economy (as the promoters of fishing, hunting, and reindeer herding), especially as weather gods. Originally the Finno-Ugric peoples probably had no concept of a hierarchic family of gods with a supreme god at its head; the attribute found in many places, "lofty" or "high," merely means "being above"—that is to say, a god appearing in the sky.

The concept of a begetting sky is stressed in southern agricultural cultures, in which an increasing importance of the Earth Mother may be observed; it is no longer a mere local field spirit but rather has the role of a great birth giver. "The god of the sky is our father, and the Earth Mother is our mother," say the Mordvins. The Earth Mother's function is not limited to field sacrifices, but also includes child giving; she is the begetter par excellence.

System of spirits. The high gods are usually encountered in connection with a rite; they are distant, invisible, and do not make surprise visits. With the guardian spirits, however, matters are different. They are first and foremost supranormal beings that appear in definite visions, auditory experiences, and other such occurrences. They appear especially when a social norm involving a guardian-spirit sanction is broken. The guardian spirits—along with the spirits of the dead—are significant as regulating factors in daily behaviour and normally are solitary local spirits, believed to "govern" and "own" a particular area: a cultural locality (*e.g.,* household spirits), a natural region (*e.g.,* guardian spirits of forest or water), or a natural element or phenomenon (*e.g.,* fire spirits or wind spirits). There are also special guardians (of man or of treasure) and various demonic beings that—though similar to the guardian spirits—are not worshiped.

The names of guardian spirits are normally compounds of words, the first element of which indicates the sphere of action and the second being a name such as "man" or "master," as in Votyak Korka-murt ("House-man") or Vu-murt ("Water-man"); "old man" or "old woman," as in Cheremis Pört-kuguza ("the Old Man of the House") or Pört-kuwa ("the Old Woman of the House"); or "father" or "mother" as in Mordvin Jurt-at'a or Jurt-ava. The system of social values is revealed by the system of guardian spirits: The house spirit protects the luck of the home; the cattle spirit watches over the cattle during the winter (in the summer the cattle come under the forest spirit); and the barn spirit looks after threshing luck. In representing these values the spirit may appear in a number of roles. Thus, the Ingrian house spirit appears as "owner," the original owner of the plot on which a house is built; "moralist," punisher of crimes against norms that may endanger the luck of the house; and "sympathizer,"



Wooden images of Ostyak house spirits.

From K.F. Karjalainen, *Die Religion der Jugra-Volker;* reproduced with permission from the Finnish Academy of Science and Letters

one who warns in advance of catastrophes threatening house or family. With some peoples—the Mordvins, for example—the guardian-spirit system is very specific and there are a very large number of spirits; with others, such as the Lapps, the Nenets, and the Ob Ugrians, there are fewer of them, and Herr der Tiere ("Master of Animals") game spirits predominate.

Sacred ancestors. The oldest form of Finno-Ugric religion is thought to be ancestor worship. Some of the main terms (*e.g.,* "grave," "hades," and "soul") go back several millennia. The cult concerned only dead members of the family; other dead beings were experienced as restless haunters, and aggressive expelling rites were used to dispel them. The worship of ancestors must be understood as a family institution in which intercourse between the living and the dead is the internal activity of a social primary group. The dead belong to the family, and they have both rights and duties; they protect the happiness of the family, assist it in its means of livelihood, and receive their share of the produce; they are also considered to be counselors, moralists, and judges. The cult of the dead can be divided in the following manner: (1) rites at the moment of death; (2) funeral preparations (washing the corpse, attiring it, and watching by it; making the coffin); (3) the committal; (4) celebrations in memory of a single dead person; (5) annual memorial ceremonies for the dead; (6) offerings and prayers to the dead in connection with earning the means of subsistence; and (7) occasional rites (*e.g.,* when moving to a new place or during illness).

The most important of the ritual ceremonies for a dead person are those that take place during the transition period, which may last for six weeks and may include addressing the departed euphemistically and in dirges. The departed person remains in the dwelling place, separated from his body; agreements are made with him about the distribution of property; he is given advice about how to live on the other side; he is invited to return for the celebration of his anniversary; and so on. The most important matter is the ensurance of harmony between the newly departed and his relations in the graveyard. Of central importance in the collective worship of the dead is the visit of the departed members to their old home; among the Eastern Finno-Ugric peoples this approximates with the Chris-

tian feast of Easter, and among the Western it is in late autumn (*e.g.,* the Finnish Kekri, November 1, an ancient festival to celebrate the seasonal change). Living members of the family also visit the graves on the anniversary days of the departed. Customs among the Lapps, the Nenets, and the northern Ostyak differ somewhat from the above; among the Lapps, the departed person is represented by a clothed log and among the Ostyak and the Nenets by a doll-effigy that is kept for as long as three years.

The otherworld is viewed as two-storied and consists of first, a graveyard hades, or underground village of the dead in a holy forest near the village; and second, a distant hades, far in the north behind the burning stream (with an admixture of paradise concepts). Name-giving rites suggest continuity and reincarnation; a child is given the name of a dead relative, and the child thereby is believed to receive the personality of the deceased relation. If the result is unsuccessful, a name-changing rite can be performed.

Divine heroes. Hero worship in Finno-Ugric religion does not point to culture heroes who are described in myth and whose actions are located in cosmogonic contexts. In general, culture heroes are not worshiped. The matter is otherwise when dealing with divinized historical figures, the cults of which are found among several of the Finno-Ugric peoples. Mardan of the Yelabuga Votyak is viewed as the progenitor of 11 villages and the one who led the dwellers therein from the north to their present habitations. There is a sacrificial ceremony in his honour every year. Also, there are signs of the worship of tribal chiefs, for example, in the forest sanctuary worship of the Votyak (*lud*) and the Volga Finns (*keremet*). The best-known of the Cheremis princes, called "the old man of the Nemda Mountain," is a great ancient warrior under whose rule the people were strong and united. According to this myth, he promised to return when war threatened; once he was called for unnecessarily and, after discovering the betrayal, he ordered the annual propitiation sacrifice of a foal. The Ob Ugrians have a large number of "local gods" of whom pictures have been made and who are sometimes associated with ancient mighty men or Christian heroes and saints. A death doll made by a shaman may also have been the origin of a hero cult; the Nenets have been known to cherish and feed such a doll for as long as 50 years.

Sacred animals. In the "hunters' religion" preserved among the northern Finno-Ugric peoples, bear ceremonies are central. The Ostyak, Vogul, Nenets, Lapps, Finns, and Karelians have all been acquainted with myths and rites connected with the bear. The myths recount that the bear is of heavenly origin and is the son of the god of the sky; it descends from heaven and, when it dies, returns there. There is also a story about a marriage between a bear and a woman from which a tribe of the Skolt Lapps (in Finland) is said to be descended. The bear-killing ceremony is divided into two acts—the killing itself and the feast afterward. Killing a bear that was protected by a forest guardian spirit involved a complicated ritual, which ended with bringing the bear home. Women believed that they had to keep at a distance so that the bear would not make them pregnant. The feast to celebrate the killing of the bear lasted two days and was full of marriage symbolism. The bear was addressed euphemistically, and a young man or woman was chosen to be its mate. A large meal made of the meat of the bear was consumed. Finally, the skull of the bear was carried in procession to the branch of a pine tree on the top of a mountain. This was the custom in Karelia. A number of miniature dramas were connected with Ob Ugrian bear rites. Masked participants tell the bear that members of a strange tribe have killed it. There seems to be a historical connection among the bear ceremonies of Ob Ugrians, Karelians, Finns, and Lapps. Nowhere else in the wide Arctic sphere have the bear songs and dramas taken such a prominent place as in this hunting ritual.

Bear ceremonies

The exogamic patrilineal clans (involving marriage outside a particular group) of the Ob Ugrians are often known by animal names—"bear," "falcon," "frog," or "dog." The animal is regarded as the manifestation of the family guardian spirit and is not allowed to be killed or eaten. Evidence of totemistic systems, in which animals are associated with blood-related groups, has been found among the Lapps and the Nenets. Some scholars consider the names of relations (animal names) found among other Finno-Ugric peoples, such as the Hungarians and Karelians, as evidence of a lost totemism.

INSTITUTIONS AND PRACTICES

Cult authorities. The male head of the family has long had a central role in leading different home and family cults. In the *lud* sanctuaries of the Votyak, for example, worship was performed by members of the family; the head of the family had the responsibility of organizing the cult and the task was hereditary. Women also were able to supervise certain minor home rituals—such as those performed in connection with cattle breeding (offerings to the guardian spirit of the cattle shed and the forest). In hunting and nomadic cultures, the head man (*e.g.,* the oldest of the hunting party or the reindeer chief) supervised the rites. The official authorities of the rites (*i.e.,* the religious specialists) among the Finno-Ugric peoples were of the following types: shamans (among the Nenets and the Lapps); seers (the counterparts of the shaman among southern peoples); sacrificing priests (the leaders of the annual rites, especially in cattle-breeding cultures and agricultural communities); guardians of the sanctuary (the protectors of holy groves, buildings, and other places and the controller of the rites); professional weeping women (the "vocalists," especially of the cult of the dead but also of weddings, who were the verbal expressers of the content of the ritual); and the masters of ceremonies at weddings. The shaman had many and various tasks in Arctic regions, but further south particular tasks were undertaken by various cult authorities: the seer (healing and counseling) and the weeping woman, or psychopomp (*i.e.,* "conductor of souls"), guiding the soul to the other world. The two last-mentioned are verbal ecstatics; the task of the seer, especially in solving critical problems, was of the utmost importance. The task of the sacrificing priest was more of a routine affair, but among the Volga Finns and the Permians for example, the long and skillful prayers as well as the complex ceremonies performed by the priests required great professional competence.

Cult centres. The home sanctuary of the Votyak is a *kuala,* a primitive log cabin near the dwelling house. In a corner at one end of the *kuala* is a shelf, at the height of a man, on which there are branches of deciduous trees and conifers, and on top of them a *voršud* (a box with a lid). A weekly offering is made here. Another Votyak sanctuary is the *lud*—a fenced-off area in an isolated place in the forest. In the middle is a primitive table for sacrificial gifts. In the *lud* regular animal sacrifices are offered and occasional crisis rites performed (sacrifices to dispel accidents or disease). The cult group in both *kuala* and *lud* is the family; the office of the sacrificing priest of the *lud* is hereditary, and in the principal house of the family there is a great *kuala,* which is visited three times a year in addition to the offerings made in the small *kuala* at home. The small *kuala* is built on a foundation of earth and ashes brought from the big *kuala.* The system is exogamous—the woman visiting the *kuala* of her own father and not that of her husband's father. The Votyak also have large groves near a spring or a brook in the vicinity of a village, where common sacrifices for the whole village are made. There are, in addition, larger sacrificing groups, which may include dozens of villages and which meet every third year for a festival lasting many weeks. The Volga Finns also have fenced-in *keremet* groves for the family cult and places of worship common to the whole village. Evidence also exists concerning sacrificial groves among the Baltic Finns and from group villages in Karelia and Ingria. In the thinly populated parts of Finland, the family cult took place either at cup stones (sacrifice stones with shallow cuplike depressions) or at holy trees. Among the nomadic Lapps (those involved in reindeer herding and fishing) *seita* ("sacrificial stone") places for worship arose near a reindeer migration route or a good fishing place, and for such a place an outstanding stone generally was chosen. The Ob Ugrians had a kind of "mobile temple"

Home and forest sanctuaries

for the wooden idols (normally kept in the corner of the house) that were placed on special sledges.

Cult practices. All the main categories of rites are found among the Finno-Ugric peoples: cyclic or calendric rites (concerning the means of livelihood), rites of passage (the transition of the individual from one status to another), and crisis rites (concerning threats of disaster). The character of these rites varies considerably, depending on ecological factors and cultural contacts. Generally, an agrarian culture produces a cult system that is more stable and formal than that produced by a mobile hunting culture or a nomadic way of life. In the latter, sacrifice rites tend to be more improvised and the cult group smaller. An example of a formal system is the distinction "upward" and "downward" in worship found among the Votyaks and the Cheremis; sacrifices of white animals are made in deciduous groves to the god of the sky and to certain nature gods, the direction of prayer being to the south; sacrifices of black animals are made to the departed and to the guardian spirits of the earth near conifers, the direction of prayer being to the north.

CONCLUSION

Two phenomena may be consistently observed with regard to the religious customs of the Finno-Ugric peoples. These are the ecological adaptation of religion and the stratification of tradition in connection with acculturation. A number of examples of the former have already been given. As far as acculturation is concerned, it may be said that the "syncretism" it produces does not result in any conflict in the religious field, except perhaps for short periods of adjustment. Old and new elements of different origins are molded into an active system, and choice and adaptation take place according to practical religious need. Christianity and Islām have in many places provided a religious superstructure, but they have not been accepted as such; certain elements from them have been adapted to the depth structure of a primitive religion. The best example of this is the preservation of folk religion in Hungary, Finland, and Estonia, where Christianity, supported by a literate culture, is ancient. Popular belief has become intertwined with the religious tradition because it has always had a function that no Christian practice has replaced. Only mass media and urbanization have jeopardized the ancient belief tradition. (L.O.H.)

Baltic religion

The term Baltic religion covers the religious beliefs and practices of the Balts, ancient inhabitants of the Baltic region of eastern Europe, who spoke languages belonging to the Baltic family of languages.

THE STUDY OF BALTIC RELIGION

Problems. The study of Baltic religion has developed as an offshoot of the study of Baltic languages—Old Prussian, Latvian, and Lithuanian (see LANGUAGES OF THE WORLD: *Indo-European languages: Baltic languages*). These form a separate group—the oldest one—of the Indo-European languages, which are closely related to the ancient Indian language Sanskrit.

Although the study of Baltic languages is important in the study of Indo-European linguistics, the study of Baltic religion has not assumed a similar level of importance in the study of comparative religion. In 1875 it was shown that the religious concepts of the Balts, when compared with those of other European peoples, are found to be marked by many older features that agree with Vedic (ancient Indian) and Iranian ideas. At least one scholarly reconstruction of ancient Indo-European religion depended mainly on Baltic religious traditions. International research in Baltic religion has, however, been greatly hindered by the fact that the languages of these small Baltic countries (Latvia and Lithuania) are but little known and because Baltic scholars have been able to work in this field only relatively recently. Thus, a comprehensive review of Baltic religion is possible only on the express understanding that many findings are only hypothetical and require further research. But, as will be seen below, even under these circumstances Baltic religious concepts help greatly in understanding the formation and structure of the oldest phases of Indo-European religion.

Age of religious concepts

Sources of data. There are four main sources of data, each with its own relevance and each requiring its own specific methodology: archaeological material, historical documents, linguistics, including toponymy (the study of the place-names of a region or language), and folklore. Since the last half of the 19th century, archaeological material has furnished much information about burial and sacrificial rites. The remains of sacred buildings have also been found. This material is of special interest in that it corroborates old religious traditions preserved by folklore, which gives added reliability to both of these sources. But archaeological material can at best furnish only a partial and incomplete picture, even though it is meaningful in some respects. Historical documents, already partially compiled and published, could be expected to yield much more information. Their value, however, is made problematic by the fact that all such documents were written by foreigners, mainly Germans who, in the course of their centuries-long eastward expansion, subjugated the Baltic peoples and exterminated some of them. Since the conquerors did not understand the Baltic languages, many documents contain the names of gods and other divinities that are without basis in fact. Baltic religion was viewed dogmatically and negatively in the light of Christian interpretations. Linguistic source material, also compiled by foreigners, shows fewer signs of interpretation, especially in regard to toponymy. Baltic folklore—one of the most extensive folklores of all European peoples—contains the greatest amount of material, especially in the form of *dainas* (short folk songs of four lines each) and folktales. Folklore is especially valuable because it contains many concepts that elsewhere have been lost under the influence of Christianity. Old religious beliefs have persisted because the Germans, after conquering the Baltic lands in the 13th and 14th centuries, made practically no attempt at Christianization and contented themselves with only economic gains. The positive result of this policy is the preservation of old traditions and religious beliefs; some researchers have also noted the similarity between the metrical structure of the *dainas* and that of the Old Indian short verses in the Rigveda (a Hindu sacred scripture).

The student of Baltic religion still encounters two difficulties. First, as has been noted, since written documents were established in Christian times, Christian influences in them are inescapable. Such influences cause difficulties and make a critical approach mandatory. Second, after the establishment of political independence of the Baltic countries following World War I, there arose a certain national romanticism that has attempted to identify Baltic culture with that of the ancient Indo-Europeans. Thus, an uncritical approach has led even to the introduction of "gods" that are actually only etymological derivations from the names of Christian saints. On the other hand, those western European scholars who are unfamiliar with the special historical and social circumstances of the Balts have assumed Baltic folklore to be on a level with the thoroughly Christianized western European folklore and thus have underestimated its importance.

MYTHOLOGY

Cosmology. In the traditions of the Baltic peoples, there are no epic myths about the creation of the world and its structure. This fact is explained by the historical and social circumstances mentioned above, which either have hindered the formation of these types of myths or, more likely, have simply made their preservation impossible. Furthermore, there has been no significant research concerning Baltic myths and their interrelationships. Fragmentary evidence found exclusively in folklore indicates only two complexes of ideas with any certainty: the first concerns the structure of the world, the second the enmity between Saule ("Sun") and Mēness (Latvian; Lithuanian Mėnulis; "Moon").

The two-region hypothesis

There is disagreement as to whether the Balts pictured the world as consisting of two regions or of three. The two-region hypothesis seems to be more plausible and is

supported by a dualism found frequently in the *dainas*: *šī saule* (literally "this sun") and *viņa saule* (literally "the other sun"). The metaphor *šī saule* symbolizes ordinary everyday human life, while *viņa saule* indicates the invisible world where the sun goes at night, which is also the abode of the dead.

The evidence does not show conclusively whether this world is located in the direction of the setting sun or under the earth, beneath which the sun travels back to the east. The sky is considered to be a mountain, sometimes of stone, and is the residence of the sky gods. Saule rides over the sky in a chariot drawn by a varying number of horses, Mēness rides to be married, and Pērkons (Latvian; Lithuanian Perkūnas; "Thunderer") makes weapons and jewelry in the sky.

The concept that Saule, unseen during the night, makes her way from west to east under the earth so that she can start her course anew over the sky mountain is also familiar. It is also possible to see here the ancient idea of a world ocean on which the earth, as a round plate, swims, an idea that has disappeared under the influence of Christianity.

The notion of a sun tree, or world tree, is one of the most important concepts regarding the cosmos. This tree grows at the edge of the path of Saule, and the setting sun (Saule) hangs her belt on the tree in preparation for rest. It is usually considered to be an oak but is also described as a linden or some other kind of tree. The tree is said to be located in the middle of the world ocean or generally to the west.

The gods. *Dievs.* The Baltic words Latvian *dievs,* Lithuanian *dievas,* and Old Prussian *deivas* are etymologically related to the Indo-European *deiu̯os;* among others, the Greek Zeus is derived from the same root. It originally meant the physical sky, but already in Old Indian and other religions the sky became personified as an anthropomorphic deity. Dievs, the pre-Christian Baltic name for God, was used by Christian missionaries (and still is) to denote the Christian God. The etymology of the word indicates that the Balts preserved its oldest forms, which is also true of the functions and attributes of the personified Baltic sky god Dievs, who lives on his farmstead on the sky mountain but does not participate in the work of the farm. Importantly, Dievs is a bridegroom who rides together with the other gods to a sky wedding in which his bride is Saule. Dievs' family is a later development; in the family, Dieva dēli ("God's Sons") play the primary role. Thus Dievs is pictured as the father of a family of sky gods. Besides such anthropomorphic characteristics, another characteristic that gives Dievs a universal significance may be observed: he appears as the creator of order in the world on the one hand, and as the judge and guardian of moral law on the other. From time to time he leaves the sky mountain and actively takes part in the everyday life of the farmers below. His participation in various yearly festivals is vividly described. In spite of this, the Baltic Dievs is similar to the Old Indian Dyaus, the Greek Zeus, and other personifications of the sky. Such divinities have a tendency, in comparison with other gods of their religions, to recede into a secondary role.

The sky god

Pērkons. In Baltic, as in other Indo-European religions, there is, in addition to Dievs, the Thunderer (Latvian Pērkons, Lithuanian Perkūnas) with quite specific functions. Pērkons is described in the oldest chronicles and in poetic and epic folklore, but, though he is a primary divinity, there is no reason to believe that he is the main god. His abode is in the sky, and, like Dievs, he sometimes descends from the sky mountain. He has two main characteristics. First, he is a mighty warrior, metaphorically described as the sky smith, and the scourge of evil. His role as adversary of the Devil and other evil spirits is of secondary importance and has been formed to a great extent under the influence of Christian syncretism. Second, he is a fertility god, and he controls the rain, an important event in the life of farmers. Various sacrifices were made to him in periods of drought as well as in times of sickness and plague. No other god occupied a place of such importance at the farmer's table during festivals, especially in the fall at harvest time. Like the other sky gods, he also has a family. Even though his daughters are mentioned occasionally, originally he had only sons, and myths depicting sky weddings portray his role vividly, as a bridegroom and as the father in his sons' weddings.

Saule. The Sun, Saule, occupies the central place in the pantheon of Baltic gods. The divinity of the sun has been recognized throughout the world, and the Balts were no exception. The Baltic description of Saule is so complete and specific that it was one of the first to be studied by scholars. Of greatest importance is the similarity in both functions and attributes of Saule and the ancient Indian god Sūrya. Similarities between the two deities are so great that, were not the two peoples separated by several thousand miles and several millennia, direct contact between them would be indicated instead of only a common origin.

The sun goddess

The representation of Saule is dualistic in that she is depicted as a mother on one hand, and a daughter on the other. Her attributes are described according to the role she plays. As a daughter she is mentioned only when she is a bride to the other sky gods. But as her daughters frequently are in the same role, it is difficult to differentiate between them. As a mother, however, she is depicted much more extensively and completely. Her farmstead on the sky mountain borders that of Dievs, and both Dieva dēli and Saules meitas ("Daughters of the Sun") play and work together. Sometimes Dievs and Saule become enraged at each other because of their respective children, as, for example, when Dieva dēli break the rings of Saules meitas or when Saules meitas shatter the swords of Dieva dēli. Their enmity lasts three days, which some scholars explain through natural phenomena; *i.e.,* the three days before the new moon when Dievs, a substitute for the moon, is not visible.

That Saule, richly described in mythology, also had a cult devoted to her is suggested by the many hymns in her honour. They contain either expressions of thanks for her bounty or prayers seeking her aid, not only in relation to agriculture but to life in general. In agriculture Saule is a sanctifier of the fertility of the fields; in the life of the individual she is a typical sky goddess, interfering in her omniscience. She has human moral characteristics and punishes the immoral and aids the suffering. Though the question of where Saule's places of worship were located is not solved, the occasions for rituals pertaining to Saule have been definitely established, the most important of which was the summer solstice. Besides song, recitative, and dance, a central place in the ceremonies was occupied by a ritual meal, at which cheese and a drink brewed with honey (later beer) were consumed.

Mēness. The Moon, Mēness, also belongs to the sky pantheon. Detailed analysis only recently has shown that he has a role as a war god in Baltic religion. Such a role is indicated not only by his dress and accoutrements but especially by his weapons and expressions used in times of war. The influence of syncretism, however, has erased the outlines of his characteristics so far as to make a description of his role and any cult he may have had very difficult. The sky wedding myths furnish a somewhat more complete picture in which he is represented as a conflict-creating rival suitor of Auseklis ("Morning Star").

The moon god

Auseklis, his sons, Dieva dēli, and Saules meitas form a separate group of divinities. Although they are mentioned in the sky myths, they have remained only as personifications of natural phenomena, characterized by the most beautiful metaphors. It is notable that a common characteristic of the sky gods, and, in fact, of all Baltic divinities, is the express tendency for each to have a family.

All of the divinities mentioned above are closely associated with horses: they either ride or are drawn in chariots across the sky mountain and arrive on earth in the same fashion. The number of horses is indeterminate but usually varies from two to five or more. This trait also confirms the close ties between Baltic and Indo-Iranian religions.

Although males form the majority of the sky gods, the chthonic (underworld) divinities are mostly female. In both Latvian and Lithuanian religions the earth is personified and called Earth Mother (Latvian Zemes māte, Lithuanian Žemyna). But the Lithuanians also have Earth Master (Žemėpatis). Latvians in general refer to mothers,

Lithuanians to masters. Zemes māte is the only deity in addition to Dievs who is originally responsible for human welfare. Based on the writings of the Roman historian Tacitus, it has been asserted that she is the mother of the other gods, but there is no support for this view in other sources. Under the influence of Christian–pagan syncretism, the Virgin Mary has assumed some of the functions of Zemes māte. Furthermore, some of these functions have been acquired and differentiated by various other later divinities, who, however, have not lost their original chthonic character. Thus, a deity of the dead has developed from Zemes māte, called in Latvian Smilšu māte ("Mother of the Sands"), Kapu māte ("Mother of the Graves"), and Veļu māte ("Mother of the Ghosts"). Libations and sacrifices were offered to Zemes māte. Such rituals were also performed in connection with the other divinities at a later stage of development. The fertility of the fields is also guaranteed by Jumis, who is symbolized by a double head of grain, and by various mothers, such as Lauka māte ("Mother of the Fields"), Linu māte ("Mother of the Flax"), and Mieža māte ("Mother of the Barley").

Forest and agricultural deities. A forest divinity, common to all Baltic peoples, is called in Latvian Meža māte and in Lithuanian Medeinė ("Mother of the Forest"). She again has been further differentiated into other divinities, or rather she was given metaphorical appellations with no mythological significance, such as Krūmu māte ("Mother of the Bushes"), Lazdu māte ("Mother of the Hazels"), Lapu māte ("Mother of the Leaves"), Ziedu māte ("Mother of the Blossoms"), and even Sēņu māte ("Mother of the Mushrooms"). Forest animals are ruled by the Lithuanian Žvėrinė opposed to the Latvian Meža māte.

The safety and welfare of the farmer's house is cared for by the Latvian Mājas gars ("Spirit of the House"; Lithuanian Kaukas), which lives in the hearth. Similarly, other farm buildings have their own patrons—Latvian Pirts māte ("Mother of the Bathhouse") and Rijas māte ("Mother of the Threshing House"); Lithuanian Gabjauja.

Gods of natural phenomena

Because natural phenomena and processes have often been raised to the level of divinities, there are a large number of beautifully described lesser mythological beings whose functions are either very limited or completely denoted by their names. Water deities are Latvian Jūras māte ("Mother of the Sea"), Ūdens māte ("Mother of the Waters"), Upes māte ("Mother of the Rivers"), and Bangu māte ("Mother of the Waves"; Lithuanian Bangpūtys), while atmospheric deities are Latvian Vēja māte ("Mother of the Wind"), Lithuanian Vėjopatis ("Master of the Wind"), Latvian Lietus māte ("Mother of the Rain"), Miglas māte ("Mother of the Fog"), and Sniega māte ("Mother of the Snow"). Even greater is the number of those beings related to human activities, but only their names are still to be found, for example Miega māte ("Mother of Sleep") and Tirgus māte ("Mother of the Market").

Goddess of destiny. Because of peculiarities of the source materials, it is difficult to determine whether the goddess of destiny, Laima (from the root word *laime,* meaning "happiness" and "luck"), originally had the same importance in Baltic religion as later, or whether her eminence is due to the specific historical circumstances of each of the Baltic peoples. In any case, a wide collection of material concerning Laima is available. The real ruler of human fate, she is mentioned frequently together with Dievs in connection with the process of creation. Although Laima determines a man's unchangeable destiny at the moment of his birth, he can still lead his life well or badly within the limits prescribed by her. She also determines the moment of a person's death, sometimes even arguing about it with Dievs.

The Devil. The Devil, Velns (Lithuanian Velnias), has a well-defined role, which is rarely documented so well in the folklore of other peoples. Besides the usual outer features, several characteristics are especially emphasized. Velns, for instance, is a stupid devil. In addition, the Balts are the only colonialized people in Europe who have preserved a large amount of folklore that in different variations and situations portrays the Devil as a German landlord. Another evil being is the Latvian Vilkacis, Lithuanian Vilkatas, who corresponds to the werewolf in the traditions of other peoples. The belief that the dead do not leave this world completely is the basis for both good and evil spirits. As good spirits the dead return to the living as invisible beings (Latvian *velis,* Lithuanian *vėlė*), but as evil ones they return as persecutors and misleaders (Latvian *vadātājs,* Lithuanian *vaidilas*).

PRACTICES, CULTS, AND INSTITUTIONS

Temples and other holy places. Archaeological excavations in the 20th century have indicated the existence of temples made of wood. The only remains of these temples are postholes. Such temples were circular, approximately 15 feet (five metres) in diameter, in the centre of which a statue of a god may have been erected. At present, however, the existence of such temples must be regarded only as conjecture within the realm of probability. On the other hand, the existence of open-air holy places or sites of worship among the Balts is confirmed by both the earliest historical documents and folklore. Such places were holy groves, called *alka* in Lithuanian. Later the word came to mean any holy place or site of worship (Lithuanian *alkvietė*). Considerable research has shown that the usual sites were little hills, where the populace gathered and sacrificed during holy festivals, all of which supports the idea that wooden buildings could have been built at these sites.

Sites of worship

Other holy places were also recognized. The most important of these appear to be bathhouses, whose function some researchers have compared to that of churches in Christianity. A large amount of evidence indicates that religious–magical rites, from birth ceremonies to funerals, were performed in such bathhouses. There are various opinions as to whether the so-called holy corner (*heilige Hinterecke*)—*i.e.,* the dark corner of a peasant's house in which a deity or patron lives—belongs to pre-Christian concepts or not. On the other hand, various places in the house proper, such as the hearth and the doorstep, were considered to be abodes of spirits. In general, the more important work sites each had its own guardian spirit. Sacrifices were performed at each spot to assure successful completion of work. Because they supplied the farmstead with water, streams and rivers were also especially important.

Religious personages. There is no reliable information that the Balts had a priestly class, let alone religious hierarchy. The 11th-century German historian Adam of Bremen, in describing conflicts between Christian missionaries and Latvians, said that "every house is filled with seers, augurers, and necromancers," which indicates that the Balts had sacral persons, probably the patriarchs of large extended families or heads of clans. As even 18th-century church inspection records show, the Christian church had great difficulty in curbing their influence, especially within their clans. Their religious functions were twofold. First, they were responsible for the welfare and means of existence of the people through the performance of appropriate rites both at work sites and during the holy festivals. Second, they assured that the proper procedure would be followed in rituals connected with the important occasions of human life, such as birth, marriage, and death. In the syncretistic amalgam of Christianity and the religion of the Balts, those persons were called sorcerers (*Zauberer*) and, according to church records, were treated by the Balts with the same reverence as bishops were treated by Christians.

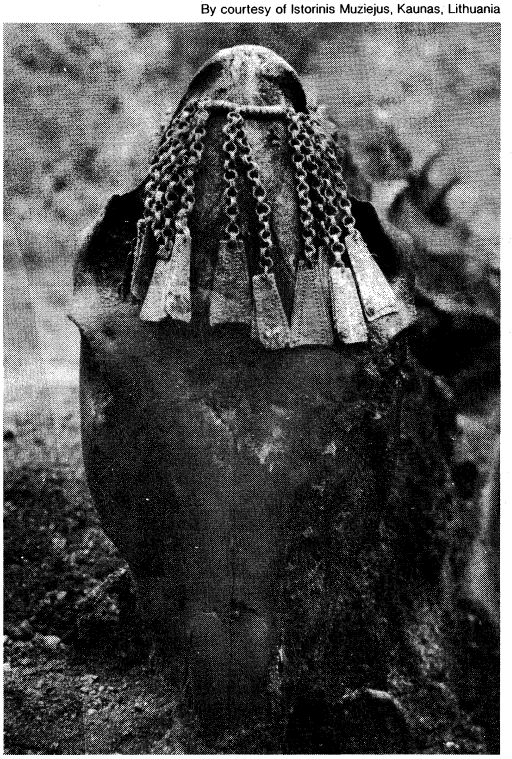
Sacred times. Special rites evolved for the festivals of the summer solstice and the harvest, while other rites were used specifically for beginning various kinds of spring work. Such spring work included sending farm animals to pasture or horses to forage for the first time, plowing the first furrow, and starting the first spring planting. The birth of a child was especially noted; it usually took place in the bathhouse or some other quiet spot. Laima was responsible for both mother and child. One birth rite, called *pirtīžas,* was a special sacral meal in which only women took part. Marriage rites were quite extensive and corresponded closely to similar Old Indian ceremonies. Fire and bread had special importance and were taken along to the house of the newly married couple. These

rites persisted until quite late and were to be seen even at the end of the 19th century, though in many cases only as games. In this connection, fire in general occupied a central place in Baltic religion. Considered holy, it was worshiped, and sacrifices were offered to it.

Death rites and customs

It seems unbelievable that even as late as 1377 and 1382, respectively, the Lithuanian king Algirdas and his brother Kęstutis could still be buried according to the old traditions in a Christian Europe; dressed in silver and gold, they were burned in funeral pyres together with their best possessions, horses, hunting dogs, birds, and weapons. In spite of a ban by the church and subsequent persecution, this rite still persisted in the 15th century. The tenacious preservation of this ancient Indo-European ritual casts light on other features of Baltic religion. Chronicles relate that Lithuanians, after losing a battle, joyfully committed suicide; this was also true of the widows of soldiers killed in battle. Such voluntary immolation and the articles buried with the dead are evidence of a belief in life after death. It is said that at the funeral of a nobleman his companions threw lynx and bear claws into the fire to aid his climb up the mountain to God, an indication of Christian influence. Archaeological excavations have also yielded evidence of fire funeral rites: the bones of humans and animals, metal jewelry, and weapons found at the sites of the funeral pyres.

By courtesy of Istorinis Muziejus, Kaunas, Lithuania



Decorated horse skull used in Baltic funeral rites. In the Kaunas State Historical Museum, Lithuania.

In funeral rites several different phases are discernible during the period between death and burning. The deceased was laid out in his house for a longer or shorter period depending on his social position and the size of his estate. During this time a meal lasting several days was held for the deceased's relatives and friends. In the course of the festivities the participants conducted fights on horseback. Lamentations, leave-takings, and praises of the deceased, as well as wishes for a safe journey to the world of the dead, accompanied the corpse on the way to the funeral pyre. In spite of persecution by the church, the tradition of lamentation has lasted until modern times, though in a somewhat modified form. One of the peculiarities of Baltic funeral rites was their similarity to wedding ceremonies. The corpse and a partner selected from the living were dressed in elaborate wedding costumes, wedding songs were sung, and dancing took place. The basis of these ceremonies was the belief that the dead anticipate a new companion with the same joy as the living do a new in-law. The corpse's living partner was a symbolical substitute for the new comrade awaited by the dead.

The use of living people to represent symbolically the companions of the dead in funerary practices suggests a dominant concept in Baltic religious thought, namely, that the boundary between the worlds of the dead and the living was not real. The dead continued to live invisibly and were present at all important occasions. A place was set for them at the festival table and no one else might sit there. The extensive practice of feeding the dead was a consequence of the concept that the living were responsible for their welfare. Originally, their food must have been placed at the hearth. In later development, meals for the dead were also placed in other buildings, such as the threshing house or the bathhouse. Under the influence of Christianity, these living dead (Latvian *velis,* Lithuanian *vėlė*) have been confused with the Devil. A widespread view was that the souls of the dead dwell in the *zalktis* (Latvian; Lithuanian *žaltys;* "green snake"); thus special care was taken in its feeding. But the *zalktis* was also closely associated with fertility and sexual symbolism.

CONCLUSION

Three main characteristics

Three main characteristics are discernible in Baltic religion. First, it is a typical astral religion in which the personified sky and main heavenly bodies play a major role. Saule, Mēness, Auseklis, and other gods have their own traits, frequently based on counterparts in nature. Although they are all related as one family, their roles within the family are varied. Depending on the cult or the plot of the myth, each divinity can assume various functions; a religious person, in general, does not experience such fluctuations as a contradiction. The second main characteristic is the personification of happiness, luck, and fate in Laima, who has assumed the role of a goddess of destiny. Because happiness is not an external, datable event, other gods besides Laima can help determine happiness in human life. The differentiation of Laima's functions has led to the establishment of some of her functions as independent entities with sometimes a poetic, sometimes a religious, meaning. The concept of destiny in Baltic religion has not, however, resulted in passive resignation or quietism but rather full exploitation of opportunities within the limits set by it. The third characteristic is the fertility cult. Here the primary force is the personified earth, called Mother, with all her functions and characteristics. It must be understood that the concept of a fertility cult entails a wider meaning, that of the assurance of human welfare in general.

These three main typological traits hardly describe Baltic religion in all of its details and nuances. The religion can also be analyzed as having two strata: one, expressed in the above three features, can be called the stable surface layer; the second, visible below the first, contains only the outlines of undifferentiated, fluid mythological and religious beings that, because of their vague character, appear in various guises and have no stable role. They are the countless house, field, and wood spirits of the nature myths.

Baltic religion, typologically, is an agricultural religion, and it is useless to speculate whether any other basis—such as nomadism, hunting, or fishing—can be found for it, because no information regarding such possibilities can be derived from any source. The amorphous agricultural clan defines the nature of Baltic religion. The farmer's gods are also farmers, though they live in great glory on their farmsteads on the sky mountain, from which they descend to help their lesser image—man. If necessary, Dievs, Saule, and Laima dress themselves in farmer's clothes and walk his fields with him. This religion does not recognize contemplation or mysticism but rather exhibits a healthy rationalism. Just as the gods are part of the cosmic order and are responsible for its maintenance, so humans obey it and become part of the divine rhythm of life set by the gods. In this way, humans cross the boundary that otherwise separates them from the world of the gods. Various specific historical circumstances explain why the Balts, in their language as well as in their religion, have preserved many elements undoubtedly belonging to the oldest phase of Indo-European religion. (H.Bi.)

Slavic religion

Slavic religion is understood here to include only the relevant beliefs and practices of the ancient Slavic peoples of eastern Europe. Slavs are usually subdivided into East Slavs (Russians, Ukrainians, and Belorussians), West Slavs (Poles, Czechs, Slovaks, and Lusatians [Sorbs]), and South Slavs (Serbs, Croats, Slovenes, Macedonians, and Bulgars).

In antiquity the Slavs were perhaps the most numerous branch of the Indo-European family of peoples. The very late date at which they came into the light of recorded history (even their name does not appear before the 6th century AD) and the scarcity of relics of their culture make serious study of them a difficult task. Sources of information about their religious beliefs are all late and by Christian hands.

SLAVIC WORLDVIEW

Socially the Slavs were organized as exogamous clans (based on marriages outside blood relationship) or, more properly, as sibs (groups of lineages with common ancestry) since marriage did not cancel membership in the clan of one's birth—a type of organization unique among Indo-European peoples. The elected chief did not have executive powers. The world had been created, in the Slavic view, once and for all, and no new law ought to modify the way of life transmitted by their ancestors. Since the social group was not homogeneous, validity and executive power were attributed only to decisions taken unanimously in an assembly, and the deliberations in each instance concerned only the question of conformity to tradition. Ancient Slavic civilization was one of the most conservative known on earth.

According to a primitive Slavic belief, a forest spirit, *leshy,* regulates and assigns prey to hunters. Its food-distributing function may be related to an archaic divinity. Though in early times the *leshy* was the protector of wild animals, in later ages it became the protector of flocks and herds. In early 20th-century Russia, if a cow or a herdsman did not come back from pasture, the spirit was offered bran and eggs to obtain a safe return.

arieties f pirits

Equally ancient is the belief in a tree spirit that enters buildings through the trunks of trees used in their construction. Every structure is thus inhabited by its particular spirit: the *domovoy* in the house, the *ovinnik* in the drying-house, the *gumenik* in the storehouse, and so on. The belief that either harmful or beneficial spirits dwell in the posts and beams of houses is still alive in the historic regions of Bosnia and Slovenia and the Poznań area of west central Poland. Old trees with fences around them are objects of veneration in Serbia and Russia and among the Slavs on the Elbe River. In 19th-century Russia a chicken was slaughtered in the drying house as a sacrifice to the *ovinnik.* This vegetal spirit is also present in the sheaf of grain kept in the "sacred corner" of the dwelling under the icon and venerated along with it, and also in noncultivated plant species that are kept in the house for propitiation or protection, such as branches of the birch tree and bunches of thistle. Such practices evidence the preagrarian origin of these beliefs. Similar to the *leshy* are the field spirit (*polevoy*), and, perhaps, the water spirit (*vodyanoy*). Akin to the *domovoy* are the spirits of the auxiliary buildings of the homestead.

MYTHOLOGY

Cosmogony. A myth known to all Slavs tells how God ordered a handful of sand to be brought up from the bottom of the sea and created the land from it. Usually, it is the Devil who brings up the sand; in only one case, in Slovenia, is it God himself. This earth-diver myth is diffused throughout practically all of Eurasia and is found in ancient India as well.

The 12th-century German missionary Helmold of Bosau recorded in *Chronica Slavorum* (*Chronicle of the Slavs*) his surprise in encountering among the Slavs on the Baltic a belief in a single heavenly God, who ignored the affairs of this world, having delegated the governance of it to certain spirits begotten by him. This is the only instance in which the sources allude to a hierarchy of divinities, but its centre is empty. The divinity mentioned by Helmold is a *deus otiosus; i.e.,* an inactive god, unique in the mythology of the Indo-European peoples. Such a deity is, however, also found among the Volga Finns, the Ugrians, and the Uralians.

Helmold's *deus otiosus*

Principal divine beings. Common to this Eurasian area is another divinity, called by Helmold and in the *Knytlinga saga* (a Danish legend that recounts the conquest of Arkona through the efforts of King Valdemar I of Denmark against the pagan and pirate Slavs) Zcerneboch (or Chernobog), the Black God, and Tiarnoglofi, the Black Head (Mind or Brain). The Black God survives in numerous Slavic curses and in a White God, whose aid is sought to obtain protection or mercy in Bulgaria, Serbia, and Pomerania. This religious dualism of white and black gods is common to practically all the peoples of Eurasia.

The Kiev Chronicle (*Povest vremennykh let*)—a 12th- to 13th-century account of events and life in the Kievan state—enumerates seven Russian pagan divinities: Perun, Volos, Khors, Dazhbog, Stribog, Simargl, and Mokosh. A Russian glossary to the 6th-century Byzantine writer John Malalas' *Chronographia* mentions a Svarog, apparently the son of Dazhbog. Of all these figures only two, Perun and Svarog, are at all likely to have been common to all the Slavs. In Polish, *piorun,* the lightning, is derived from the name of Perun, and not vice versa. In the province of Wielkopolska the expression *do pierona*—meaning "go to the Devil"—has been recorded. In the expression, *pieron/piorun* is no longer the lightning but the being who launches it. Uncertain or indirect traces of Perun are also encountered among the Carpathians and in Slovenia and Serbia. The lightning-wielding Perun cannot be considered the supreme god of the Slavs but is rather a spirit to whom was given the governance of the lightning.

In Estonia the prophet Elijah is considered to be the successor to Ukko, the ancient spirit of lightning. Similarly, the prophet Elijah replaces Elwa in Georgia and Zeus in Greece. It is therefore probable that, among the Slavs also, Elijah is to be considered a successor of Perun. According to a popular Serbian tradition, God gave the lightning to Elijah when he decided to retire from governing the world. The Serbian story agrees with Helmold's description of the distribution of offices by an inactive God. Elijah is a severe and peevish saint. It is rare that his feast day passes without some ill fortune. Fires—even spontaneous combustion—are blamed on him.

A similar complex may be seen if the Slavic Perun is equated with Perkūnas, the lightning deity of the Lithuanians. In Latvia, creatures with black fur or plumage were sacrificed to Pērkons, as they were to the fire god Agni in ancient India. Such deities are therefore generic deities of fire, not specifically celestial and even less to be regarded as supreme. Scholarly efforts to place Perun at the centre of Slavic religion and to create around him a pantheon of deities of the Greco-Roman type cannot yield appreciable results. Russian sources treat Svarog, present as Zuarasici among the Liutici of Rethra (an ancient locality in eastern Germany), as a god of the drying-house fire. But the Belorussians of Chernigov, when lighting the drying-house fire, invoke Perun and not Svarog, as if Svarog (apparently from *svar,* "litigation" or "dispute," perhaps referring to the friction between the pieces of wood used to produce ignition) were an appellation of Perun.

Folk conceptions. In a series of Belorussian songs a divine figure enters the homes of the peasants in four forms in order to bring them abundance. These forms are: *bog* ("god"); *sporysh,* anciently an edible herb, today a stalk of grain with two ears, a symbol of abundance; *ray* ("paradise"); and *dobro* ("the good"). The word *bog* is an Indo-Iranian word signifying riches, abundance, and good fortune. *Sporysh* symbolizes the same concept. In Iranian *ray* has a similar meaning, which it probably also had in Slavic languages before it acquired the Christian meaning of paradise. *Bog,* meaning "riches," connotes grain. The same concept is also present in Mordvinian *pa* and *riz*—where their provenance is certainly Iranian. Among the Mordvins, Paz, like the Slavic Bog, enters into the homes bringing abundance. The adoption of the foreign word *bog* probably displaced from the Slavic languages the Indo-

European name of the celestial God, Deivos (Ancient Indian Deva, Latin Deus, Old High German Ziu, etc.), which Lithuanian, on the other hand, has conserved as Dievas.

Worship of celestial bodies

Among the heavenly bodies the primary object of Slavic veneration was the moon. The name of the moon is of masculine gender in Slavic languages (Russian *mesyats;* compare to Latin *mensis*). The word for sun (Russian *solntse*), on the other hand, is a neuter diminutive that may derive from an ancient feminine form. In many Russian folk songs a verb having the sun as its subject is put in the feminine form, and the sun is almost always thought of as a bride or a maiden.

It is to the moon that recourse is had to obtain abundance and health. The moon is saluted with round dances and is prayed to for the health of children. During lunar eclipses, weapons are discharged at the monsters who are said to be devouring the moon, and weeping and wailing express the sharing of the moon's sufferings. In Serbia the people have always envisioned the moon as a human being. Such appellations as father and grandfather are customarily applied to the moon in Russian, Serbo-Croatian, and Bulgarian folk songs. At Risano (modern-day Risan, Yugos.) in the days of the 19th-century writer Vuk Karadžić—the father of modern Serbian literature—it was said of a baby four months old that he had four grandfathers. In Bulgaria the old people teach small children to call the moon Dedo Bozhe, Dedo Gospod ("Uncle God, Uncle Lord"). Ukrainian peasants in the Carpathians openly affirm that the moon is their god and that no other being could fulfill such functions if they were to be deprived of the moon. In two Great Russian supplications the sickle moon is invoked as "Adam"—the final phase of a fully developed moon worship in which the moon becomes the progenitor of the human family.

PRACTICES, CULTS, AND INSTITUTIONS

Places of worship. Though the idols of which the Russian chronicles speak appear to have been erected out-of-doors, the German chronicles provide detailed descriptions of enclosed sacred places and temples among the Baltic Slavs. Such enclosures were walled and did not differ from profane fortifications—areas usually of triangular shape at the confluence of two rivers, fortified with earthwork and palisades, especially on the access side. The fortifications intended for religious purposes contained wooden structures including a cell for the statue of a god, also made of wood and sometimes covered in metal. These representations, all anthropomorphic, very often had supernumerary bodily parts: seven arms, three or five heads (Trigelavus, Suantevitus, and Porenutius). The temples were in the custody of priests, who enjoyed prestige and authority even in the eyes of the chiefs and received tribute and shares of military booty. Human sacrifices, including eviscerations, decapitations, and trepanning, had a propitiatory role in securing abundance and victory. One enclosure might contain up to four temples; those at Szczecin (Stettin), in northwestern Poland, were erected in close proximity to each other. They were visited annually by the whole population of the surrounding district, who brought with them oxen and sheep destined to be butchered. The boiled meat of the animals was distributed to all the participants without regard to sex or age. Dances and plays, sometimes humorous, enlivened the festival.

Temple ceremonies

Communal banquets and related practices. The custom of communal banquets has been preserved into modern times in Russia in the *bratchina* (from *brat,* "brother"), in the *mol'ba* ("entreaty" or "supplication"), and in the *kanun* (a short religious service); in the Serbian *slava* ("glorification"); and in the *sobor* ("assembly") and *kurban* ("victim" or "prey") of Bulgaria. Formerly, communal banquets were also held by the Poles and the Polabs (Elbe Slavs) of Hannover. In Russia the love feasts are dedicated to the memory of a deceased person or to the patron saint of the village and in Serbia to the protecting saint from whom the *rod* or *pleme* ("clan") took its name. Scholars no longer have any doubts of the pre-Christian nature of these banquets. The Serbian *slava* is clearly dedicated to a saint held to be the founder of the clan. These saints are patrons or founders and are all men who have died. When the Serbs celebrate the *slava* of the prophet Elijah or of the archangel Michael they do not set out the "dead man's plate" (the *koljivo,* boiled wheat), because Elijah and Michael are not dead. In certain localities in Serbia even the women given in marriage to another clan, the so-called *odive,* have to be present at the *slava.* They return with their children (according to the ancient matrilineal conception of the offspring), but not with their husbands, who belong to another clan and celebrate another *slava.* More akin to the ancient pagan feasts of the Baltic is the Serbian *seoska slava,* or "*slava* of the village," in which the whole population of the place takes part and consumes in common the flesh of the victims prepared in the open air. Such feasts are votive. In Russia sometimes the animals (or their flesh) are first brought into the church and perfumed with incense. Even at the beginning of the 20th century there were small villages in Russia where cattle were butchered only on the occasion of these festivities, three or four times a year. *The Homily of Opatoviz* (attributed to Herman, bishop of Prague) of the 10th–11th century emphatically condemns the love feasts as well as the veneration of statues and Slavic worship of the dead and veneration of saints as if they were gods. As in the Christian era the saints entered the line of ancestors, so perhaps in pagan antiquity ancient divinities (Perun, Svarog) were taken over as tribal progenitors. The Slavs did not record genealogies, and the founders of their clans were mainly legendary. The social unit sought to assure for itself the favour of powerful figures of the past, even of more than one, representing them in several forms on the same pillar or giving to their statues supernumerary bodily parts that would express their superhuman powers. A hollow bronze idol, probably ancient Russian, was found at Ryazan, Russian S.F.S.R. The idol has four faces with a fifth face on its breast.

The eastern Finns and the Ugrians venerated their dead in the same way, similarly representing them as polycephalic (multiple-headed), and also held communal banquets in their honour. Wooden buildings (the so-called *continae*) in which the faithful Baltic Slavs used to assemble for amusement, to deliberate, or to cook food have been observed in the 20th century among the Votyaks, the Cheremis, and the Mordvins, but especially among the Votyaks. Such wooden buildings also existed sparsely in Slavic territory in the 19th century, in Russia, particularly in the Ukraine, as well as here and there among the South Slavs.

If it is supposed that, as among the Finns and the Ugrians, each clan venerated its own divine ancestor in a separate building, this would explain why many sacred enclosures would contain more than one *contina*—three at Carentia (the island of Garz at the mouth of the Oder River) and four at Szczecin.

The system of idolatry of the Baltic area was essentially manistic (pertaining to worship of ancestors). It is not irrelevant that until the 19th century there survived here and there throughout the Danubian-Balkan region the custom of reopening graves three, five, or seven years after interment, taking out the bones of the corpses, washing them, wrapping them in new linen, and reinterring them. Detailed descriptions of this procedure have come particularly from Macedonia and Slovenia. Among East and West Slavs only faint echoes of the custom of a second interment survive in folk songs. In the former *guberniya* (province) of Vladimir, east of Moscow, as late as 1914, when a grave was to be dug, a piece of cloth was taken along with which to wrap the bones of any earlier corpse that might be unearthed in the process of digging. Such corpses would then be reinterred with the newly deceased. In protohistoric times the tumuli (mounds) of the mortuaries of the Krivichi, a populous tribe of the East Slavs of the northwest, the so-called long kurgans (burial mounds), contained cinerary urns buried in the tumulus together and all at one time. Such a practice could occur only as the consequence of collective and simultaneous cremation. There must, therefore, have existed a periodic cremation season or date, as for the opening of the tombs in Macedonia and as has been verified elsewhere in comparing the South Asian areas of second interment, in preparation for which the corpses are temporarily ex-

The custom of second interment

humed. The cremations by the Krivichi are of exhumed bones. In the Volga region today the Mordvins still burn the disinterred bones of the dead in the flames of a "living fire" ignited by friction.

Considering the religious past of the Slavs, it is not surprising that manism was strong enough to epitomize and overwhelm all or practically all of their religious views. The seasonal festivals of the Slavs turn out to be almost entirely dedicated to the dead, very often without the participants realizing it, as in the case of the Koljada (Latin Kalendae)—the annual visit made by the spirits of the dead, under the disguise of beggars, to all the houses in the village. It is possible that the bones of the disinterred were kept for a long period inside the dwellings, as is still sometimes done in the Tyrol of Austria, and that the sacred corner—now occupied by the icon—was the place where they were kept.

The spirits of the departed are not only venerated but also feared, especially the spirits of those who were prematurely deprived of life and its joys. It is believed that such spirits are greedy for the good things thus lost and that they make attempts to return to life—to the peril of the living. They are the prematurely dead, the so-called unclean dead. Particularly feared are maidens who died before marriage and are believed to be addicted to the kidnapping of bridegrooms and babies. One annual festival in particular, the Semik (seventh Thursday after Easter) was dedicated to the expulsion of these spirits. They are called *rusalki* in Russia, *vile* or *samovile* in Serbo-Croatia and Bulgaria.

The dead person who does not decompose in the grave becomes a vampire, a word and concept of Slavic origin. To save the living from a vampire's evil deeds, it is necessary to plant a stake in the grave so that it passes through the heart of the corpse or else to exhume the corpse and burn it. Since the classes of unclean dead are believed to have been constantly increasing (in Macedonia, for example, it is believed that all those born in the three months between Christmas and Lady Day are unclean), then all of the dead—once objects of veneration and piety—will at some point be in danger of rancor, fear, and eventual disregard. A Christian clergy that has lent its presence at the exhumation and destruction of vampires has thereby contributed unwittingly to the preservation of this last phase of Slavic paganism into modern times.

There are other rites associated with second interment of which the Slavs have forgotten the purpose, such as the cemetery pyres—fires lit on top of the tombs—or the assiduous watering of graves. In Polynesia and South America where second interment is practiced, these same acts have the purpose of fostering decomposition of the corpses in order to hasten exhumation.

Numerous other ritual acts are performed by the Slavs, for the most part related to this complex of beliefs. In 19th-century Russia, if a man encountered the procession of naked women who were plowing a furrow around the village at night in order to protect it from an epidemic, he was inevitably killed. It was a chthonic (underworld) being to which, in those same times, human sacrifices were offered in Russia (more rarely in Poland and Bulgaria), since the victims were often buried alive. In most cases they were either voluntary victims or chosen by lot from among the devotees. Since such acts were punished by the law of the state, the sacrifices were performed in secrecy and are difficult to document. (E.G.)

Greek religion

Greek religion, comprising the beliefs of the ancient Hellenes about gods and their relationship with humanity, lasted in its developed form for more than a thousand years, from the time of Homer (probably 9th or 8th century BC) to the reign of the emperor Julian (4th century AD), though its origins may be traced to the remotest eras. During that period its influence spread as far west as Spain, east to the Indus River, and throughout the Mediterranean world. Its effect was most marked on the Romans, who identified their deities with the Greek. Under Christianity, Greek heroes and even deities survived as saints, while the rival madonnas of southern European communities reflected the independence of local cults. The rediscovery of Greek literature during the Renaissance and, above all, the novel perfection of classical sculpture produced a revolution in taste that had far-reaching effects on Christian religious art. The most striking characteristic of Greek religion was the belief in a multiplicity of anthropomorphic deities, coupled with a minimum of dogmatism.

The student of Greek religion is naturally concerned to know what the Greeks believed about their gods. They had numerous beliefs, but the sole requirement was to believe that the gods existed and to perform ritual and sacrifice, through which the gods received their due. To deny the existence of a deity was to risk reprisals, from the deity or from other mortals. The list of avowed atheists is brief. But if a Greek went through the motions of piety, he risked little, since no attempt was made to enforce orthodoxy, a religious concept almost incomprehensible to the Greeks. The Greeks had no word for religion itself, the closest approximations being *eusebeia* ("piety") and *threskeia* ("cult"). The large corpus of myths concerned with gods, heroes, and rituals embodied the worldview of Greek religion and remains its legacy. It should be noted that the myths varied over time and that, within limits, a writer—*e.g.*, a Greek tragedian—could vary a myth in order to change not only the role played by the gods in it but also the evaluation of the gods' actions. From the later 6th century BC onward, myths and gods were subject to rational criticism on ethical or other grounds. In these circumstances it is easy to overlook the fact that most Greeks "believed" in their gods in roughly the modern sense of the term and that they prayed in a time of crisis not merely to the "relevant" deity but to any deity on whose aid they had established a claim by sacrifice. To this end, each Greek polis had a series of public festivals throughout the year that were intended to ensure the aid of all the gods who were thus honoured. They reminded the gods of services rendered and asked for a quid pro quo. In crises in particular the Greeks, like the Romans, were often willing to add deities borrowed from other cultures.

It is frequently difficult to obtain evidence of Greek religious practice, not only within the mystery cults but also more generally. In the latter case, the reason is not one of secrecy; the Greeks simply did not anticipate a posterity that would be different from themselves. Religious practices were universally known—as were such everyday activities as sailing triremes and holding assemblies—and it was not deemed necessary to record these things. It should be remembered that Pausanias, the most important source for a number of topics, was writing in the 2nd century AD, and that even by the 5th century BC the meaning and origins of some of the practices he described were evidently unknown.

HISTORY

The roots of Greek religion. The study of a religion's history includes the study of the history of those who espoused it, together with their spiritual, ethical, political, and intellectual experiences. Greek religion as it is currently understood probably resulted from the mingling of religious beliefs and practices between the incoming Greek-speaking peoples who arrived from the north during the 2nd millennium BC and the indigenous inhabitants whom they called "Pelasgi." The incomers' pantheon was headed by the Indo-European sky god variously known as Zeus (Greek), Dyaus (Indian), or Jupiter (Roman Diespater). But there was also a Cretan sky god, whose birth and death were celebrated in rituals and myths quite different from those of the incomers. The incomers applied the name of Zeus to his Cretan counterpart. In addition, there was a tendency, fostered but not necessarily originated by Homer and Hesiod, for major Greek deities to be given a home on Mount Olympus. Once established there in a conspicuous position, the Olympians came to be identified with local deities and to be assigned as consorts to the local god or goddess. An unintended consequence (since the Greeks were monogamous) was that Zeus in particular became markedly polygamous. (Zeus already had a consort when he arrived in the Greek world and

Greek and Pelasgian deities

took Hera, herself a major goddess in Argos, as another.) Hesiod used—or sometimes invented—the family links among the deities, traced out over several generations, to explain the origin and present condition of the universe. At some date, Zeus and other deities were identified locally with heroes and heroines from the Homeric poems and called by such names as Zeus Agamemnon. The Pelasgian and the Greek strands of the religion of the Greeks can sometimes be disentangled, but the view held by some scholars that any belief related to fertility must be Pelasgian, on the grounds that the Pelasgi were agriculturalists while the Greeks were nomadic pastoralists and warriors, seems somewhat simplistic. Pastoralists and warriors certainly require fertility in their herds—not to mention in their own number. In cult, Athena, a warrior goddess and patron of the arts and crafts and a prominent Olympian, presided also over fertility festivals. The citizens prayed to her for all good things; the fertility of field, flock, and citizen was as essential to the well-being of the polis as its victory in war.

The Archaic period. Sometime before the Homeric poems took their present form, the orgiastic cult of the nature divinity Dionysus reached Greece, traditionally from Thrace and Phrygia. Because the god's name is Greek, it has been suggested that his worship represents not a novelty but a reversion to Mycenaean religion. His devotees, armed with *thyrsoi* (wands tipped with a pine cone and wreathed with vine or ivy) and known as maenads (literally "mad women"), were reputed to wander in *thiasoi* (revel bands) about mountain slopes, such as Cithaeron or Parnassus; the practice persisted into Roman imperial times. They were also supposed, in their ecstasy, to practice the *sparagmos,* the tearing of living victims to pieces and feasting on their raw flesh (*ōmophagia*). While such behaviour continued in the wild, in the cities—in Athens, at any rate—the cult of Dionysus was tamed before 500 BC. Tragedy developed from the choral song of Dionysus.

The cult of Dionysus

In the 7th and 6th centuries BC "tyrants" (monarchs whose position was not derived from heredity) seized power in many poleis. Some of them, such as Peisistratus in Athens, were nobles themselves and rose to power by offering the poor defense against the rest of the nobility. Once established, Peisistratus built temples and founded or revived festivals. At this time, too, the earliest references to the Eleusinian Mysteries appear. The Mysteries offered a more personal, less distant relationship with the divine than did most of the Olympians. There was no Eleusinian way of life. On one or two occasions (depending on the grade they wished to attain) the initiates went to Eleusis; what they saw there in the place of initiation sufficed to ensure them a life after death that was much more "real" than the Olympian belief that the dead were witless ghosts.

The Classical period. During the 6th century BC the rationalist thinking of Ionian philosophers had offered a serious challenge to traditional religion. At the beginning of the 5th century, Heracleitus of Ephesus and Xenophanes of Colophon heaped scorn on cult and gods alike.

The Sophists, with their relentless probing of accepted values, continued the process. Little is known of the general success of these attacks in society as a whole. The Parthenon and other Athenian temples of the late 5th century proclaim the taste and power of the Athenians rather than their awe of the gods; but it is said that after the completion of Phidias' chryselephantine Athena on the Acropolis, the old olive-wood statue of Athena, aesthetically no match for Phidias' work, continued to receive the worship of most of the citizens. Antiquity evoked awe; some of the most revered objects in Greece were antique and aniconic figures that bore the name of an Olympian deity.

Festivals were expressive of religion's social aspect and attracted large gatherings (*panēgyreis*). Mainly agrarian in origin, they were seasonal in character, held often at full moon and on the 7th of the month in the case of Apollo, and always with a sacrifice in view. Many were older than the deity they honoured, like the Hyacinthia and Carneia in Laconia, which were transferred from local heroes to Apollo. The games were a special festival, sometimes part of other religious events. Some festivals of Athens were performed on behalf of the polis and all its members. Many of these seem to have been originally the cults of individual noble families who came together at the *synoikismos,* the creation of the polis of Athens from its small towns and villages. The nobles continued to furnish the priests for these cults, but there was, and could be, no priestly class. There were no "priests of the gods," or even priests of an individual god; one became a priest of one god at one temple. Except for these public festivals, anyone might perform a sacrifice at any time. The priest's role was to keep the temple clean; he was usually guaranteed some part of the animal sacrificed. A priesthood offered a reasonably secure living to its incumbent.

Popular religion flourished alongside the civic cults. Peasants worshiped the omnipresent deities of the countryside, such as the Arcadian goat-god Pan, who prospered the flocks, and the nymphs (who, like Eileithyia, aided women in childbirth) inhabiting caves, springs (Naiads), trees (Dryads and Hamadryads), and the sea (Nereids). They also believed in nature spirits such as Satyrs and Sileni and equine Centaurs. Among the more popular festivals were the rural Dionysia, which included a phallus pole; the Anthesteria, when new wine was broached and offerings were made to the dead; the Thalysia, a harvest celebration; the Thargelia, when a scapegoat (*pharmakos*) assumed the communal guilt; and the Pyanepsia, a bean feast in which boys collected offerings to hang on the *eiresiōne* ("wool pole"). Women celebrated the Thesmophoria in honour of Demeter and commemorated the passing of Adonis with laments and miniature gardens, while images were swung from trees at the Aiora to get rid of an ancient hanging curse. Magic was widespread. Spells were inscribed on lead tablets. Statues of Hecate, goddess of witchcraft, stood outside dwellings, while Pan's image was beaten with herbs in time of meat shortage.

The Hellenistic period. Greek religion, having no creed, did not proselytize. In the heyday of the polis, the Greek religion was spread by the founding of new poleis, whose colonists took with them part of the sacred fire from the hearth of the mother city and the cults of the city's gods. ("Heroes," being essentially bound to the territory in which they were buried, had to be left behind.) There was a tendency for Greeks to identify the gods of others with their own, often at a superficial level. So the virgin Artemis was identified with the chief goddess of Ephesus, a fertility deity. After Alexander the Great had created a political world in which the poleis were engulfed by large kingdoms, those deities who were not too closely linked with a particular place became more prominent. Mystery cults, which offered a personal value to the individual in a large and indifferent world, also flourished. The Cabeiri of Samothrace were patronized by both the Ptolemies and the Romans, while the Egyptian cults of Isis and Sarapis, in a Hellenized form, spread widely. Rulers sometimes officially invited new gods to settle in times of crisis, in the hope that they would strive on their new worshipers' behalf against their mortal foes: a mode of religious thought that flourished at least until the days of the Roman emperor Constantine. Those novel cults that seemed likely to pose a threat to public order, on the other hand, were suppressed by the Romans. The Senate destroyed the Bacchic cult in Italy in 186 BC for the same reasons as Trajan gave to Pliny for his treatment of the Christians: Any cult in which men and women, bond and free, could participate and meet together—a most unusual circumstance in the ancient world—had dangerous political implications.

BELIEFS, PRACTICES, AND INSTITUTIONS

The gods. The early Greeks personalized every aspect of their world, natural and cultural, and their experiences in it. The earth, the sea, the mountains, the rivers, custom-law (*themis*), and one's share in society and its goods were all seen in personal as well as naturalistic terms. When Achilles fights with the River in the *Iliad,* the River speaks to Achilles but uses against him only such weapons as are appropriate to a stream of water. In Hesiod, what could be distinguished as anthropomorphic deities and personalizations of natural or cultural phenomena both beget and are begotten by each other. Hera is of the first type—god-

dess of marriage but not identified with marriage. Earth is evidently of the second type, as are, in a somewhat different sense, Eros and Aphrodite (god and goddess of sexual desire) and Ares (god of war). These latter are personalized and anthropomorphized, but their worshipers may be "filled" with them. Some deities have epithets that express a particular aspect of their activities. Zeus is known as Zeus Xenios in his role as guarantor of guests. It is possible that Xenios was originally an independent deity, absorbed by Zeus as a result of the Olympocentric tendencies of Greek religion encouraged by the poems of Homer and Hesiod.

In Homer the gods constitute essentially a super-aristocracy. The worshipers of these gods do not believe in reward or punishment after death; one's due must come in this life. Every success shows that the gods are well disposed, for the time being at least; every failure shows that some god is angry, usually as a result of a slight, intended or unintended, rather than from the just or unjust behaviour of one mortal to another. The Greeks knew what angered their mortal aristocracy and extrapolated from there. Prayer and sacrifice, however abundant, could not guarantee that the gods would grant success. The gods might prefer peace on Olympus to helping their worshipers. These are not merely literary fictions; they reflect the beliefs of people who knew that though it might be necessary to offer prayer and sacrifice to the gods, it was not sufficient. Greek and Trojans sacrificed to their gods to ensure divine support in war and at other times of crisis. It was believed that Zeus, the strongest of the gods, had favoured the Trojans, while Hera had favoured the Greeks. Yet Troy fell, like many another city. The Homeric poems here offer an explanation for something that the Greek audience might at any time experience themselves.

There is no universal determinism in Homer or in other early writers. *Moira* ("share") denotes one's earthly portion, all the attributes, possessions, goods, or ills that together define one's position in society. Homeric society is stratified, from Zeus to the meanest beggar. To behave in accordance with one's share is to behave in accordance with one's status; and even a beggar may go beyond his share, though he is likely to be punished for it. Zeus, the most powerful entity in Homer's universe, certainly has the power to go beyond his share; but if he does so, the other gods "will not approve." And Zeus may be restrained, unless he feels that his "excellence," his ability to perform the action, is being called into question. Then he may insist on displaying his excellence, as do Achilles and Agamemnon, whose values coincide with those of Zeus in such matters.

tratifica-
ion of
ociety

In Homer, *hērōs* denotes the greatest of the living warriors. The cults of these mighty men developed later around their tombs. Heroes were worshiped as the most powerful of the dead, who were able, if they wished, to help the inhabitants of the polis in which their bones were buried. Thus, the Spartans brought back the bones of Orestes from Tegea. Historical characters might be elevated to the status of heroes at their deaths. During the Peloponnesian War, the inhabitants of Amphipolis heroized the Spartan general Brasidas, who had fought so well and bravely and died in their defense. It is power, not righteousness, that distinguishes the hero; it is the feeling of awe before the old, blind Oedipus that stimulates the Thebans and the Athenians to quarrel over his place of burial. Since they are the mightiest of the dead, heroes receive offerings suitable for chthonic deities.

Cosmogony. Of several competing cosmogonies in archaic Greece, Hesiod's *Theogony* is the only one that has survived in more than fragments. It records the generations of the gods from Chaos (literally, "Yawning Gap") through Zeus and his contemporaries to the gods who had two divine parents (*e.g.,* Apollo and Artemis, born of Zeus and Leto) and the mortals who had one divine parent (*e.g.,* Heracles, born of Zeus and Alcmene). Hesiod uses the relationships of the deities, by birth, marriage, or treaty, to explain why the world is as it is and why Zeus, the third supreme deity of the Greeks, has succeeded in maintaining his supremacy—thus far—where his predecessors failed. Essentially, Zeus is a better politician and has the balance of power, practical wisdom, and good counsel on his side. (Whether Hesiod or some earlier thinker produced this complex nexus of relationships, with which Hesiod could account for virtually anything that had occurred or might occur in the future, the grandeur of this intellectual achievement should not be overlooked.)

Man. In the period in Greece between Homer and about 450 BC the language of relationships between god and god, man and god, and human beings of lower status with human beings of higher status was the same. The deities remained a super-aristocracy. There was a scale of "power-and-excellence" on which the position of every human being and every deity could be plotted. Both god and man were likely to resent any attempt of an inferior to move higher on the scale. It constituted *hubris* ("overweening pride") for a Greek *heros* to claim that he would have a safe voyage whether or not the gods were willing; it was likewise *hubris* for Electra to presume to criticize the behaviour of her mother Clytemnestra.

A further reason for Olympian disapproval, only marginally present in Homer, was the pollution caused by certain actions and experiences, such as childbirth, death, or having a bad dream. The divine world of the Greeks was bisected by a horizontal line. Above that line were the Olympians, gods of life, daylight, and the bright sky; below it were the chthonic (underworld) gods of the dead and of the mysterious fertility of the earth. The Olympians kept aloof from the underworld gods and from those who should be in their realm: Creon is punished in Sophocles' *Antigone* by the Olympians for burying Antigone alive, for she is still "theirs," and for failing to bury the dead Polyneices, gobbets of whose flesh are polluting their altars; Hippolytus is abandoned by Artemis, her most ardent worshiper, as his death approaches, for all corpses pollute. Pollution was not a moral concept; and it further complicated relationships between the Greeks and their gods.

Eschatology. In Homer only the gods were by nature immortal, but Elysium was reserved for their favoured sons-in-law, who they exempted from death. Heracles alone gained a place on Olympus by his own efforts. The ordinary hero hated death, for the dead were regarded as strengthless doubles who had to be revived with drafts of blood, mead, wine, and water in order to enable them to speak. They were conducted, it was believed, to the realm of Hades by Hermes; but the way was barred, according to popular accounts, by the marshy river Styx. Across this, Charon ferried all who had received at least token burial, and coins were placed in the mouths of corpses to pay the fare. Originally only great sinners like Ixion, Sisyphus, and Tityus, who had offended the gods personally, were punished in Tartarus. But the doctrines of the Orphics influenced Pindar, Empedocles, and, above all, Plato. According to the latter, the dead were judged in a meadow by Aeacus, Minos, and Rhadamanthus and were consigned either to Tartarus or to the Isles of the Blest. Long periods of purgation were required before the wicked could regain their celestial state, while some were condemned forever. The dead were permitted to choose lots for their next incarnation. Subsequently they drank from the stream of Lethe, the river of oblivion, and forgot all of their previous experiences.

Immortality and death

Sacred writings. Greek religion was not based on a written creed or body of dogma. Nevertheless, certain sacred writings survive in the form of hymns, oracles, inscriptions, and instructions to the dead. Most elaborate are the Homeric Hymns, some of which may have been composed for religious festivals, though their subject matter is almost entirely mythological. Delphic inscriptions include hymns to Apollo but, like the Epidaurian hymn by Isyllus to Asclepius, they are not concerned with liturgy. Delphic oracles are quoted from literary sources but appear, on the whole, to be retrospective concoctions, like the Hebraic–Hellenistic collection of Sibylline prophecies. Questions scratched on folded lead tablets have been found at Dodona, and detailed instructions to the dead, inscribed on gold leaf and possibly of Orphic inspiration, have been found in Greek graves in southern Italy. Papyrus fragments of similar character have been recovered from graves in Macedonia and Thessaly.

Shrines and temples. In the earliest times deities were worshiped in awesome places such as groves, caves, or mountain tops. Mycenaean deities shared the king's palace. Fundamental was the precinct (temenos) allotted to the deity, containing the altar, temple (if any), and other sacral or natural features, such as the sacred olive in the temenos of Pandrosos on the Athenian Acropolis. *Naoi* (temples—literally "dwellings"—that housed the god's image) were already known in Homeric times and, like models discovered at Perachora, were of wood and simple design. Poros and marble replaced wood by the end of the 7th century BC, when temples became large and were constructed with rows of columns on all sides. The image, crude and wooden at first, was placed in the central chamber (cella), which was open at the eastern end. No ritual was associated with the image itself, though it was sometimes paraded. Hero shrines were far less elaborate and had pits for offerings. Miniature shrines also were known.

Oracles and divination

Most oracular shrines included a subterranean chamber, but no trace of such has been found at Delphi, though the Pythia was always said to "descend." At the oracle of Trophonius, discovered in 1967 at Levádhia, incubation was practiced in a hole. The most famous centre of incubation was that of Asclepius at Epidaurus. His temple was furnished with a hall where the sick were advised by the demigod in dreams. Divination was also widely practiced in Greece. Augurs interpreted the flight of birds, while dreams, and even sneezes, were regarded as ominous. Seers also divined from the shape of altar smoke and the conformation of victims' entrails.

Priesthood. Even in the state cults, priesthoods were frequently ancestral prerogatives. Eteobutads organized the cult of the hero-king Erechtheus at Athens; Praxiergids superintended the washing of Athena's robes at the Plynteria; Clytiads and Iamids officiated at the altar of Zeus at Olympia. Although there was no official clergy, since the religious and secular spheres were not sharply divided, professional assistance was available at sacrifices. There was no necessary correspondence between the sex of deities and that of priests. Hera and Athena favoured priestesses, but Isis and Cybele favoured priests. Apollo again inspired the Pythia (priestess) at Delphi but a priest at Ptoon. The mysteries at Eleusis were administered by the Eumolpids and Kerykes. The latter assembled the initiates (*mystae*), while the former provided the Hierophant, who revealed the mysteries in the torchlit Anaktoron (king's shrine) within the great Telesterion, or entrance hall.

Festivals. The precise details of many festivals are obscure. Among the more elaborate was the Panathenaea, which was celebrated at high summer, and every fourth year (the Great Panathenaea) on a more splendid scale. Its purpose, besides offering sacrifice, was to provide the ancient wooden image of Athena, housed in the "Old Temple," with a new robe woven by the wives of Athenian citizens. The Great Panathenaea included a procession, a torch race, athletic contests, mock fights, and bardic recitations. The Great Dionysia was celebrated at Athens in spring. At the end of the ritual the god's image was escorted to the theatre of Dionysus, where it presided over the dramatic contests. It, like its rural counterpart, included phallic features.

André Held—Ziolo

Painted Greek vase showing a Dionysiac feast, 450–425 BC. In the Louvre, Paris.

Olympic Games

The Olympic Games formed part of the great festival of Zeus held every fourth summer in the god's sacred precinct—the Altis beside the river Alpheius in the western Peloponnese. A truce was proclaimed in order to permit any warring Greeks to compete, and the celebrations lasted five days. Sacrifice and libation were made at the altar of Zeus, where omens were taken and oracles proclaimed, and at the tomb of Pelops and the altar of Hestia. Competitors and judges took the oath to observe the rules, processions were held, bards recited, and winners were honoured at state banquets. The richer and more famous were immortalized by lyric poets, such as Simonides, Bacchylides, and Pindar. Though women were banned, girls competed at the festival of Hera. The games held in honour of Zeus at Nemea, Apollo at Delphi, and Poseidon at the Isthmus followed the Olympian pattern.

Rites. Sacrifice was offered to the Olympian deities at dawn at the altar in the temenos, which normally stood east of the temple. Representing as it did a gift to the gods, sacrifice constituted the principal proof of piety. The gods were content with the burnt portion of the offering, while the priests and worshipers shared the remainder of the meat. Different animals were sacred to different deities—*e.g.,* heifers to Athena, cows to Hera, pigs to Demeter, bulls to Zeus and Dionysus, dogs to Hecate, game and heifers to Artemis, horses to Poseidon, and asses to Priapus—though the distinctions were not rigorously observed. The practices of ritual washing before sacrifice, sprinkling barley grains, and making token offerings of hair are described by Homer. Victims were required to be free of blemish, or they were likely to offend the deity. Sacrifice also was made to chthonian powers in the evening. Black victims were offered, placed in pits, and the meat was entirely consumed. Sacrifice preceded battles, treaties, or similar events. Human sacrifice appears, if it was practiced at all, to have been the exception. Bloodless sacrifices were made to some deities and heroes.

Prayers and processions

Prayers normally began with compliments to the deity, followed by discreet references to the petitioner's piety, and ended with his special plea. In addressing a prayer to an Olympian, the suppliant stood with his arms raised palm upward. Processions formed part of most gatherings (*panēgyreis*) and festivals. The Panathenaic procession set out from the Pompeion (sacred storehouse) at dawn, headed by maiden basket-bearers (*kanēphoroi*), who carried the sacred panoply. Elders bore boughs (*thallophoroi*) while youths (*ephēboi*) conducted the victims for sacrifice, and cavalry brought up the rear. The robe was spread on the mast of a wheeled ship.

The procession to Eleusis to restore the sacred objects, brought by the *ephēboi* to the Eleusinium some time previously, followed the wooden image of Iacchus (a personification of the ritual cry), which was escorted by its own priest, the *iacchagogos,* and officials. The *mystae* wore myrtle crowns and carried sheaves of grain. Whatever the nature of the mysteries, those initiated returned in a mood of exaltation. Adepts (*epoptai*) were later admitted to more solemn rites (to see an ear of wheat, scoffers said).

Religious art and iconography. Art often portrays incidents relevant to the study of Greek religion, but frequently essential information is missing. On a well-known sarcophagus from Ayías Triádhos in Crete, for example, a priestess dressed in a skin skirt assists at a sacrifice, flanked by wreathed axes on which squat birds. The significance of the scene has been much discussed. The birds have been regarded as epiphanies of deities, giving sacral meaning to the transformations in Homer. Again, since goddesses appear to preponderate in Minoan–Mycenaean art, while male deities are represented on an inferior scale, this has been thought to reflect the general superiority of goddesses in many parts of Greece. In the earliest period, terra-cotta statuettes of deities were small and crude, while the old cult images were made of wood and commonly attributed to Daedalus. When artists turned to bronze and marble,

Painting showing a dead man (right) receiving an offering of a votive ship and two calves, and the priestess (left) pouring a libation. From a terra-cotta sarcophagus from the necropolis of Ayías Triádhos, Crete, *c.* 1400 BC. In the Archaeological Museum, Crete.
Gad Borel—Boissonnas

they depicted the anthropomorphic deities as idealized human beings. The skill of the Greek sculptor reached an almost unparalleled height in the new temples on the Acropolis of Athens; but while high attainment in the visual arts indicates the presence of a high level of aesthetic consciousness, it would be hazardous to conclude that it necessarily accompanied a profound religious experience. The human form idealized was still used for portraying the gods, but only a brief step was needed to produce an art in which the human form was idealized for its own sake. The growth and decline of religions may be matched by the growth and decline of their art, and works of high artistic quality may inspire, and be inspired by, profound religious emotions; but, as the continued worship of the old wooden aniconic statue of Athena, mentioned above, indicates, it is often the antiquity of a cult object that inspires the awe that surrounds it.

Apart from cult statues and dedications like the Acropolis *korai* ("maidens"), the gods frequently were represented on the pediments, metopes, and friezes of temples, usually in mythological scenes. For the details of ritual, vase painting has proved a fruitful source of information. Dionysiac subjects are common, though usually imaginary, but cult scenes and fertility customs also appear.

If "Greek religion" is understood to denote the beliefs about the Greek gods and their relationships with humanity as recorded in surviving writings from the Homeric poems onward, Greek religion was always evolving. Cultic activity, however, was conservative, as it is in most cultures. Practices continued to be observed that were no longer understood by the worshipers. High claims have been made, and continue to be made, for the quality of Greek religion as a religion, with ethical deities and strong tendencies toward monotheism. Indeed, this is probably the orthodox view. Those who contest it hold that it is incautious to extrapolate from a few scattered passages in a Greek author to produce a systematic theology that can then be used to interpret the rest of the work under discussion. The debate shows no sign of coming to an end; but the heterodox are wont to observe that Xenophanes, Pindar, and Plato evidently read Greek literature in the same way as the heterodox propose that it should be read. Plato's strictures in Books II and III of *The Republic* and elsewhere on Greek religion as he knew it bear eloquent testimony to this.

MYTHOLOGY

mpor-
ance of
Greek
nythology
n the
Vestern
vorld

Although people of all countries, eras, and stages of civilization have developed myths that explain the existence and workings of natural phenomena, recount the deeds of gods or heroes, or seek to justify social or political institutions, the myths of the Greeks have remained unrivaled in the Western world as sources of imaginative and appealing ideas. Poets and artists from ancient times to the present have derived inspiration from Greek mythology and have discovered contemporary significance and relevance in classical mythological themes.

Sources of myths: literary and archaeological. *The Homeric poems: the Iliad and the Odyssey.* Herodotus remarked that Homer and Hesiod gave to the Olympian gods their familiar characteristics. Few today would accept this literally. In the first book of the *Iliad,* the son of Zeus and Leto (Apollo, line 9) is as instantly identifiable by his patronymic as are the sons of Atreus (Agamemnon and Menelaus, line 16). In both cases, the audience is expected to have knowledge of the myths that preceded their literary rendering. Most scholars hold that Homer's tone is light and humorous and that the audience is not expected to take the gods seriously. Others reply that little is known to suggest that the Greeks treated Homer, or any other source of Greek myths, as mere entertainment, whereas there are prominent Greeks from Pindar to the later Stoa for whom myths, and those from Homer in particular, are so serious as to warrant bowdlerization or allegorization.

Assumption of prior knowledge

The works of Hesiod: Theogony and Works and Days. The fullest and most important source of myths about the origin of the gods is the *Theogony* of Hesiod. The elaborate genealogies mentioned above are accompanied by folktales and etiological myths. The *Works and Days* shares some of these in the context of a farmer's calendar and an extensive harangue on the subject of justice addressed to Hesiod's possibly fictitious brother Perses. The orthodox view treats the two poems as quite different in theme and treats the *Works and Days* as a theodicy (a natural theology). It is possible, however, to treat the two poems as a diptych, each part dependent on the other. The *Theogony* declares the identities and alliances of the gods, while the *Works and Days* gives advice on the best way to succeed in a dangerous world rendered yet more dangerous by its gods; and Hesiod urges that the most reliable—though by no means certain—way is to be just.

Other literary works. Fragmentary post-Homeric epics, of varying date and authorship, filled the gaps in the accounts of the Trojan War recorded in the *Iliad* and *Odyssey;* the so-called Homeric Hymns (shorter surviving poems) are the source of several important religious myths. Many of the lyric poets preserved various myths, but the odes of Pindar of Thebes (flourished 6th–5th century BC) are particularly rich in myth and legend. The works of the three tragedians—Aeschylus, Sophocles, and Euripides, all of the 5th century BC—are remarkable for the variety of the traditions they preserve. In Hellenistic times (323–30 BC) Callimachus, a 3rd-century-BC poet and scholar in Alexandria, recorded many obscure myths; his contemporary, the mythographer Euhemerus, suggested that the gods were originally human, a view known as Euhemerism. Apollonius of Rhodes, another scholar of the 3rd century BC, preserved the fullest account of the Argonauts in search of the Golden Fleece. In the period of the Roman Empire, the *Library* of the pseudo-Apollodorus (attributed to a 2nd-century-AD scholar), the antiquarian writings of the Greek biographer Plutarch, and the works of Pausanias, a 2nd-century-AD geographer, as well as the *Genealogies* of Hyginus, a 2nd-century-AD mythographer, have provided valuable sources in Latin of later Greek mythology.

Archaeological discoveries. The discovery of the Mycenaean civilization by Heinrich Schliemann, a 19th-century German amateur archaeologist, and the discovery of the Minoan civilization in Crete (from which the Mycenaean ultimately derived) by Sir Arthur Evans, a 20th-century English archaeologist, helped to explain many of the ques-

Significance of Mycenaean and Minoan archaeological discoveries

tions about Homer's epics and provided archaeological proofs of many of the mythological details about gods and heroes. Unfortunately, the evidence about myth and ritual at Mycenaean and Minoan sites is entirely monumental, because the Linear B script (an ancient form of Greek found in both Crete and Greece) was mainly used to record inventories, though the names of gods and heroes have been doubtfully revealed.

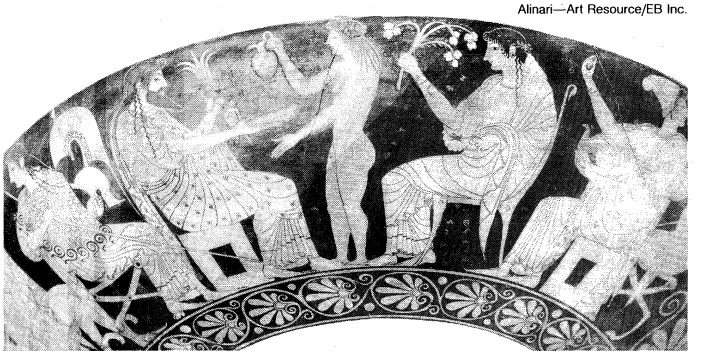
Geometric designs on pottery of the 8th century BC depict scenes from the Trojan cycle, as well as the adventures of Heracles. The extreme formality of the style, however, renders much of the identification difficult, and there is no inscriptional evidence accompanying the designs to assist scholars in identification and interpretation. In the succeeding Archaic (*c.* 750–*c.* 500 BC), Classical (*c.* 480–323 BC), and Hellenistic periods, Homeric and various other mythological scenes appear to supplement the existing literary evidence.

Forms of myth in Greek culture. To distinguish among myth, legend, and folktale can be useful, provided it is remembered that the Greeks themselves did not do so.

Religious myths. Greek religious myths are concerned with gods or heroes in their more serious aspects or are connected with ritual. They include cosmogonical tales of the genesis of the gods and the world out of Chaos, the successions of divine rulers, and the internecine struggles that culminated in the supremacy of Zeus, the ruling god of Olympus. They also include the long tale of Zeus's amours with goddesses and mortal women, which usually resulted in the births of younger deities and heroes. The goddess Athena's unique status is implicit in the story of her motherless birth (she was born directly from Zeus); and the myths of Apollo explain that god's sacral associations, describe his remarkable victories over monsters and giants, and stress his jealousy and the dangers inherent in immortal alliances.

Myths of Dionysus, on the other hand, demonstrate the hostility aroused by a novel faith. Some myths are closely associated with rituals, such as the account of the drowning of the infant Zeus's cries by the Curetes, attendants of Zeus, clashing their weapons, or Hera's annual restoration of her virginity by bathing in the spring Canathus. Some myths about heroes and heroines also had a religious basis. The tale of man's creation and moral decline forms part of the myth of the Four Ages (see below). His subsequent destruction by flood and regeneration from stones is partly based on folktale.

Alinari—Art Resource/EB Inc.



The gods on Olympus: Athena, Zeus, Dionysus, Hera, and Aphrodite; detail of a painting on a Greek cup. In the Museo Municipale, Tarquinia, Italy.

Themes of legends and folktales

Legends. Myths were viewed as embodying divine or timeless truths, whereas legends (or sagas) were quasihistorical. Hence, famous events in epics, such as the Trojan War, were generally regarded as having really happened, and heroes and heroines were believed to have actually lived. Earlier sagas, such as the voyage of the Argonauts, were accepted in a similar fashion. Most Greek legends were embellished with folktales and fiction, but some certainly contain a historical substratum. Such are the tales of more than one sack of Troy, which are supported by archaeological evidence, and the labours of Heracles, which suggest Mycenaean feudalism. Again, the legend of the Minotaur (a being part human, part bull) could have arisen from exaggerated accounts of bull leaping in ancient Crete.

In another class of legends, heinous offenses, such as attempting to make love to a goddess against her will, deceiving the gods grossly by inculpating them in crime, or assuming their prerogatives, were punished by everlasting torture in the underworld. The consequences of social crimes, such as murder or incest, were also described in legend (*e.g.,* the story of Oedipus, who killed his father and married his mother). Legends were also sometimes employed to justify existing political systems or to bolster territorial claims.

Folktales. Folktales, consisting of popular recurring themes and told for amusement, inevitably found their way into Greek myth. Such is the theme of lost persons—whether husband, wife, or child (*e.g.,* Odysseus, Helen of Troy, or Paris of Troy)—found or recovered after long and exciting adventures. Journeys to the land of the dead were made by Orpheus (a hero who went to Hades to restore his dead wife, Eurydice, to the realm of the living), Heracles, Odysseus, and Theseus (the slayer of the Minotaur). The victory of the little man by means of cunning against impossible odds, the exploits of the superman (*e.g.,* Heracles), or the long-delayed victory over enemies are still as popular with modern writers as they were with the Greeks. The successful countering of the machinations of cruel sires and stepmothers (who are often witches), rescues of princesses from monsters, or temporary forgetfulness at a crucial moment are also familiar themes in Greek myth. Recognition by tokens, such as Odysseus' scar or peculiarities of dress, is another common folktale motif. The babes-in-the-wood theme of the exposure of children and their subsequent recovery is also found in Greek myth. The Greeks, however, also knew of the exposure of children as a common practice.

Types of myths in Greek culture. *Myths of origin.* Myths of origin represent an attempt to render the universe comprehensible in human terms. Greek creation myths (cosmogonies) and views of the universe (cosmologies) were more systematic and specific than those of other ancient peoples. Yet their very artistry serves as an impediment to interpretation, since the Greeks embellished the myths with folktale and fiction told for its own sake. Thus, though the aim of Hesiod's *Theogony* is to describe the ascendancy of Zeus (and, incidentally, the rise of the other gods), the inclusion of such familiar themes as the hostility between the generations, the enigma of woman (Pandora), the exploits of the friendly trickster (Prometheus), or struggles against powerful beings or monsters like the Titans (and, in later tradition, the Giants) enhances the interest of an epic account.

According to Hesiod, four primary divine beings first came into existence: the Gap (Chaos), Earth (Gaea), the Abyss (Tartarus), and Love (Eros). The creative process began with the forcible separation of Gaea from her doting consort Heaven (Uranus) in order to allow her progeny to be born. The means of separation employed, the cutting off of Uranus' genitals by his son Cronus, bears a certain resemblance to a similar story recorded in Babylonian epic. The crudity is relieved, however, in characteristic Greek fashion by the friendly collaboration of Uranus and Gaea, after their divorce, in a plan to save Zeus from the same Cronus, his cannibalistic sire.

According to Greek cosmological concepts, the Earth was viewed as a flat disk afloat on the river of Ocean. The Sun (Helios) traversed the heavens like a charioteer and sailed around the Earth in a golden bowl at night. Natural fissures were popularly regarded as entrances to the subterranean house of Hades, home of the dead.

Myths of the ages of the world. From a very early period, Greek myths seem open to criticism and alteration on grounds of morality or of misrepresentation of known facts. In the *Works and Days,* Hesiod makes use of a scheme of Four Ages (or Races): Golden, Silver, Bronze, and Iron. "Race" is the more accurate translation, but "Golden Age" has become so established in English that both terms should be mentioned. These races or ages are separate creations of the gods, the Golden Age belonging to the reign of Cronus, the subsequent races the creation

The Four Ages of Hesiod

of Zeus. Those of the Golden Age never grew old, were free from toil, and passed their time in jollity and feasting. When they died, they became guardian spirits on Earth.

Why the Golden Age came to an end Hesiod failed to explain, but it was succeeded by the Silver Age. After an inordinately prolonged childhood, the men of the Silver Age began to act presumptuously and neglected the gods. Consequently, Zeus hid them in the Earth, where they became spirits among the dead.

Zeus next created the men of the Bronze Age, men of violence who perished by mutual destruction. At this point the poet intercalates the Age (or Race) of Heroes. He thereby destroys the symmetry of the myth, in the interests of history: what is now known as the Minoan–Mycenaean period was generally believed in antiquity to have been a good time to live. (This subjection of myth to history is not universal in Greece, but it is found in writers such as Hesiod, Xenophanes, Pindar, Aeschylus, and Plato.) Of these heroes the more favoured (who were related to the gods) reverted to a kind of restored Golden Age existence under the rule of Cronus (forced into honourable exile by his son Zeus) in the Isles of the Blessed.

The final age, the antithesis of the Golden Age, was the Iron Age, during which the poet himself had the misfortune to live. But even that was not the worst, for he believed that a time would come when infants would be born old, and there would be no recourse left against the universal moral decline. The presence of evil was explained by Pandora's rash action in opening the fatal urn.

Elsewhere in Greek and Roman literature, the belief in successive periods or races is found with the belief that by some means, when the worst is reached, the system gradually (Plato, *Politikos*) or quickly (Virgil, *Fourth Eclogue*) returns to the Golden Age. Hesiod may have known this version; he wishes to have been born either earlier or later. There is also a myth of progress, associated with Prometheus, god of craftsmen; but the progress is limited, for the 19th-century concept of eternal advancement is absent from Greek thought.

Myths of the gods. Myths about the gods described their births, victories over monsters or rivals, love affairs, special powers, or connections with a cultic site or ritual. As these powers tended to be wide, the myths of many gods were correspondingly complex. Thus, the Homeric Hymns to Demeter, a goddess of agriculture, and to the Delian and Pythian Apollo describe how these deities came to be associated with sites at Eleusis, Delos, and Delphi, respectively. Similarly, myths about Athena, the patroness of Athens, tend to emphasize the goddess' love of war and her affection for heroes and the city of Athens; and those concerning Hermes (the messenger of the gods), Aphrodite (goddess of love), or Dionysus describe Hermes' proclivities as a god of thieves, Aphrodite's lovemaking, and Dionysus' association with wine, frenzy, miracles, and even ritual death. Poseidon (god of the sea) was unusually atavistic, in that his union with Earth and his equine adventures appear to hark back to his pre-marine status as a horse or earthquake god. Many myths are treated as trivial and lighthearted; but, as was said above, this judgment rests on the suppressed premise that any divine behaviour that seems inappropriate for a major religion must have seemed absurd and fictitious to the Greeks. It is uncertain whether Homer knew of the judgment of Paris; but he knew the far from trivial consequences for Troy of the favour of Aphrodite and the bitter enmity of Hera and Athena, which the judgment of Paris was composed to explain.

Combinations of myths, legends, and folktales

As time went on, an accretion of minor myths continued to supplement the older and more authentic ones. Thus, the loves of Apollo, virtually ignored by Homer and Hesiod, explained why the bay (or laurel) became Apollo's sacred tree and how he came to father Asclepius, a healing god. Similarly, the presence of the cuckoo on Hera's sceptre at Hermione or the invention of the panpipe were explained by fables. Such etiological myths proliferated during the Hellenistic era, though in the earlier periods genuine examples are harder to detect.

Of folk deities, the nymphs (nature goddesses) personified nature or the life in water or trees and were said to punish unfaithful lovers. Water nymphs (Naiads) were reputed to drown those with whom they fell in love, such as Hylas, a companion of Heracles. Even the gentle Muses (goddesses of the arts and sciences) blinded their human rivals, such as the bard Thamyris. Satyrs (youthful folk deities with bestial features) and Sileni (old and drunken folk deities) were the nymphs' male counterparts. Like sea deities, Sileni possessed secret knowledge that they would reveal only under duress. Charon, the grisly ferryman of the dead, was also a popular figure of folktale.

Myths of heroes. Hero myths included elements from tradition, folktale, and fiction. The saga of the Argonauts, for example, is highly complex and includes elements from folktale and fiction, but the information that the fleet mustered at Colchis may be regarded as genuine legend. Episodes in the Trojan cycle, such as the departure of the Greek fleet from Aulis or Theseus' Cretan expedition and death on Scyros, may belong to traditions dating from the Minoan–Mycenaean world. On the other hand, events described in the *Iliad* probably owe far more to Homer's creative ability than to genuine tradition. Even heroes like Achilles, Hector, or Diomedes are largely fictional, though doubtlessly based on legendary prototypes. The *Odyssey* is the prime example of the wholesale importation of folktales into epic. All the best-known Greek hero myths, such as the labours of Heracles and the adventures of Perseus, Cadmus, Pelops, or Oedipus, depend more for their interest on folktales than legend. Certain heroes—Heracles, the Dioscuri (the twins Castor and Pollux), Amphiaraus (one of the Argonauts), or Hyacinthus (a youth loved by Apollo and accidentally killed)—may be regarded as partly legend and partly religious myth. Thus, whereas Heracles, a man of Tiryns, may originally have been a historical character, the myth of his demise on Oeta and subsequent elevation to full divinity is closely linked with a cult. In time, Heracles' popularity was responsible for connecting his story with the Argonauts, an earlier attack on Troy, and with Theban myth. Similarly, the exploits of the Dioscuri are those of typical heroes: fighting, carrying off women, and cattle rustling. After their death they passed six months alternately beneath the Earth and in the world above, which suggests that their worship, like that of Persephone (the daughter of Zeus and Demeter), was connected with fertility or seasonal change.

Myths of seasonal renewal. Certain myths, in which goddesses or heroes were temporarily incarcerated in the underworld, were allegories of seasonal renewal. Perhaps the best-known myth of this type is the one telling how Hades (Latin Pluto), the god of the underworld, carried Persephone off to be his consort, causing her mother Demeter, the goddess of grain, to allow the earth to grow barren out of grief. Because of her mother's grief, Zeus permitted Persephone to spend four months of the year in the house of Hades and eight in the light of day. In less benign climates, she was said to spend six months of the year in each. Some scholars hold that Persephone's time below ground represents the summer months, when Greek fields are parched and bare; but the *Hymn to Demeter,* the earliest source, states explicitly that Persephone returns when the spring flowers are flourishing (line 401). Myths of seasonal renewal, in which the deity dies and returns to life at particular times of the year, are plentiful. An important Greek example is the Cretan Zeus, mentioned above.

Myths involving theriolatry. Many Greek myths involve animal transformations, though there is no proof that theriolatry (animal worship) was ever practiced by the Greeks. Gods sometimes assumed the form of beasts in order to deceive goddesses or women. Zeus, for example, assumed the form of a bull when he carried off Europa, a Phoenician princess, and appeared in the guise of a swan in order to attract Leda, wife of a king of Sparta. Poseidon took the shape of a stallion to beget the wonder horses Arion and Pegasus.

Deities appearing in nonhuman form

These myths do not suggest theriolatry. No worship is offered to the deity concerned. The animals serve other purposes in the narratives. Bulls were the most powerful animals known to the Greeks and may have been worshiped in the remote past. But for the Greeks in even the earliest sources, there is no indication that Zeus or

Heracles fighting with the Amazons, detail from a volute crater attributed to Euphronius, *c.* 500 BC. In the Museo Archeologico, Arezzo, Italy.

Alinari—Art Resource/EB Inc.

Poseidon were once bulls or horses, or that Hera was ever "ox-eyed" other than metaphorically, or that "gray-eyed" Athena was ever "owl-faced."

Deities appearing in human form

Other types. Other types of myth exemplified the belief that the gods sometimes appeared on Earth disguised as men and women and rewarded any help or hospitality offered them. Baucis, an old Phrygian woman, and Philemon, her husband, for example, were saved from the flood by offering hospitality to Zeus and Hermes, both of whom were in human form. The punishment of men's presumption in claiming to be the gods' superiors, whether in musical skill or even the number of their children, is described in several myths. The gods' jealousy of their musical talents appears in the beating and flaying of the flute-playing Satyr, Marsyas, by Athena and Apollo, as well as in the attaching of ass's ears to King Midas for failing to appreciate the superiority of Apollo's music to that of the god Pan. Jealousy was the motive for the slaying of Niobe's many children, because of Niobe's flaunting her fecundity to the goddess Leto, who had only two offspring. Similar to such stories are the moral tales about the fate of Icarus, who flew too high on homemade wings, or the myth about Phaethon, the son of Helios, who failed to perform a task too great for him (controlling the horses of the Sun).

Transformation into flowers or trees, whether to escape a god's embraces (such as Daphne, a nymph transformed into a laurel tree), as the result of an accident (such as Hyacinthus, a friend of Apollo, who was changed into a flower), or because of pride (*e.g.,* the beautiful youth Narcissus who fell in love with his own reflection and was changed into a flower), were familiar themes in Greek myth.

Also popular were myths of fairylands, such as the Garden of the Hesperides (in the far west) or the land of the Hyperboreans (in the far north), or encounters with monstrous or outlandish people, such as the Centaurs or Amazons.

Greek mythological characters and motifs in art and literature. People of all eras have been moved and baffled by the deceptive simplicity of Greek myths, and Greek mythology has had a profound effect on the development of Western civilization.

The earliest visual representations of mythological characters and motifs occur in late Mycenaean and sub-Mycenaean art. Though identification is controversial, Centaurs, a Siren, and even Zeus's lover Europa have been recognized. Mythological and epic themes are also found in Geometric art of the 8th century BC, but not until the 7th century did such themes become popular in both ceramic and sculptured works. During the Classical and subsequent periods, they became commonplace. The birth of Athena was the subject of the east pediment of the Parthenon in Athens, and the legend of Pelops and the labours of Heracles was the subject of the corresponding pediment and the metopes (a space on a Doric frieze) of the Temple of Zeus at Olympia. The battles of gods with Giants and of Lapiths (a wild race in northern Greece) with Centaurs were also favourite motifs. Pompeian frescoes reveal realistic representations of Theseus and Ariadne, Perseus, the fall of Icarus, and the death of Pyramus.

Resurgence of Greek mythological motifs

The great Renaissance masters added a new dimension to Greek mythology. Among the best-known subjects of Italian artists are Botticelli's "Birth of Venus," the Ledas of Leonardo da Vinci and Michelangelo, and Raphael's "Galatea."

Through the medium of Latin and, above all, the works of Ovid, Greek myth influenced medieval poets such as Petrarch and Boccaccio in Italy and Chaucer in England; Dante in Italy during the Renaissance; and, later, the English Elizabethans and John Milton. Racine in France and Goethe in Germany revived Greek drama, and nearly all the major English poets from Shakespeare to Robert Bridges turned for inspiration to Greek mythology. In more recent times, classical themes have been reinterpreted by such major dramatists as Jean Anouilh, Jean Cocteau, and Jean Giraudoux in France, Eugene O'Neill in America, and T.S. Eliot in England and by great novelists such as James Joyce (Irish) and André Gide (French). The German composers Christoph Gluck (18th century) and Richard Strauss (20th century), the German-French composer Jacques Offenbach (19th century), and many others have set Greek mythological themes to music.

(J.R.T.P./A.W.H.A.)

Roman religion

This section deals with the beliefs and practices of the inhabitants of the Italian peninsula from ancient times until the ascendancy of Christianity in the 4th century AD.

NATURE AND SIGNIFICANCE

The Romans, according to the orator and politician Cicero, excelled all other peoples in the unique wisdom that made them realize that everything is subordinate to the rule and direction of the gods. Yet Roman religion was based not on divine grace but instead on mutual trust (*fides*) between god and man. The object of Roman religion was to secure the cooperation, benevolence, and "peace" of the gods (*pax deorum*). The Romans believed that this divine help would make it possible for them to master the unknown forces around them that inspired awe and anxiety (*religio*), and thus they would be able to live successfully. Consequently, there arose a body of rules,

Object of Roman religion

the *jus divinum* ("divine law"), ordaining what had to be done or avoided.

These precepts for many centuries contained scarcely any moral element; they consisted of directions for the correct performance of ritual. Roman religion laid almost exclusive emphasis on cult acts, endowing them with all the sanctity of patriotic tradition. Roman ceremonial was so obsessively meticulous and conservative that, if the various partisan accretions that grew upon it throughout the years can be eliminated, remnants of very early thought can be detected near the surface.

This demonstrates one of the many differences between Roman religion and Greek religion, in which such remnants tend to be deeply concealed. The Greeks, when they first began to document themselves, had already gone quite a long way toward sophisticated, abstract, and sometimes daring conceptions of divinity and its relation to man. But the orderly, legalistic, and relatively inarticulate Romans never quite gave up their old practices. Moreover, until the vivid pictorial imagination of the Greeks began to influence them, they lacked the Greek taste for seeing their deities in personalized human form and endowing them with mythology. In a sense, there is no Roman mythology, or scarcely any. Although discoveries in the 20th century, notably in the ancient region of Etruria (between the Tiber and Arno rivers, west and south of the Apennines), confirm that Italians were not entirely unmythological, their mythology is sparse. What is found at Rome is chiefly only a pseudomythology (which, in due course, clothed their own nationalistic or family legends in mythical dress borrowed from the Greeks). Nor did Roman religion have a creed; provided that a Roman performed the right religious actions, he was free to think what he liked about the gods. And, having no creed, he usually deprecated emotion as out of place in acts of worship.

In spite, however, of the antique features not far from the surface, it is difficult to reconstruct the history and evolution of Roman religion. The principal literary sources, antiquarians such as the 1st-century-BC Roman scholars Varro and Verrius Flaccus, and the poets who were their contemporaries (under the late Republic and Augustus), wrote 700 and 800 years after the beginnings of Rome. They wrote at a time when the introduction of Greek methods and myths had made erroneous (and flattering) interpretations of the distant Roman past unavoidable. In order to supplement such conjectures or facts as they may provide, scholars rely on surviving copies of the religious calendar and on other inscriptions. There is also a rich, though frequently cryptic, treasure-house of material in coins and medallions and in works of art.

HISTORY

Early Roman religion. For the earliest times, there are the various finds and findings of archaeology. But they are not sufficient to enable scholars to reconstruct archaic Roman religion. They do, however, suggest that early in the 1st millennium BC, though not necessarily at the time of the traditional date for the founding of Rome (753 BC), Latin and Sabine shepherds and farmers with light plows came from the Alban Hills and the Sabine Hills, and that they proceeded to establish villages at Rome, the Latins on the Palatine Hill and the Sabines (though this is uncertain) on the Quirinal and Esquiline hills. About 620 the communities merged, and *c.* 575 the Forum Romanum between them became the town's meeting place and market.

Deification of functions. From such evidence it appears that the early Romans, like many other Italians, sometimes saw divine force, or divinity, operating in pure function and act, such as in human activities like opening doors or giving birth to children, and in nonhuman phenomena such as the movements of the sun and seasons of the soil. They directed this feeling of veneration both toward happenings that affected human beings regularly and, sometimes, toward single, unique manifestations, such as a mysterious voice that once spoke and saved them in a crisis (Aius Locutius). They multiplied functional deities of this kind to an extraordinary degree of "religious atomism," in which countless powers or forces were identified with one phase of life or another. Their functions were sharply defined; and in approaching them it was important to use their right names and titles. If one knew the name, one could secure a hearing. Failing that, it was often best to cover every contingency by admitting that the divinity was "unknown" or adding the precautionary phrase "or whatever name you want to be called" or "if it be a god or goddess."

Veneration of objects. The same sort of anxious awe was extended not only to functions and acts but also to certain objects that inspired a similar belief that they were in some way more than natural. This feeling was aroused, for example, by springs and woods, objects of gratitude in the torrid summer, or by stones that were often believed to be meteorites—*i.e.,* had apparently reached the earth in an uncanny fashion. To these were added products of human action, such as burial places and boundary stones, and inexplicable things, such as Neolithic implements (probably the mysterious meteorites were often these) or bronze shields (artifacts that had strayed in from more advanced cultures).

To describe the powers in these objects and functions that inspired the *horror,* or sacred thrill, the Romans eventually employed the word *numen,* suggestive of a god's nod, *nutus;* though so far there is no evidence that this usage was earlier than the 2nd century BC. The application of the word spirit to *numen* is anachronistic in regard to early epochs because it presupposes a society capable of greater abstraction. Nor must the term mana, used by Melanesians to describe their own concept of superhuman forces, be introduced too readily. The two societies are not necessarily analogous and, besides, the deduction from such comparisons that the Romans experienced an impersonal, pre-deistic, primordial stage of religion that neatly preceded the personal stage cannot be regarded as correct. On the contrary, from the very earliest times, the supernatural forces that they envisaged included a number of deities in analogous human forms; among them were certain "high gods." Foremost among these was a divinity of the sky, Jupiter, akin to the sky gods of other early Indo-European-speaking peoples, the Sanskrit Dyaus and Greek Zeus. Not yet, probably, a Supreme Being, though superior in some sense to other divine powers, this god of the heavens was easily linked with the forces of function and object, with lightning and weather, or with the uncanny stone that came from on high and was called Jupiter Lapis. High gods

Purpose of sacrifice and magic. These gods and sacred functions and objects seemed charged with power because they were mysterious and alarming. In order to secure their food supply, physical protection, and growth in numbers, the early Romans believed that such forces had to be propitiated and made allies. Sacrifice was necessary. The product sacrificed would revitalize the divinity, which was seen as a power of action and therefore likely to run down unless so revitalized. By this nourishment he or it would become able and ready to fulfill requests. And so the sacrifice was accompanied by the phrase *macte esto!* ("be you increased!").

Prayer was a normal accompaniment of sacrifice, and as a conception of the divine powers gradually developed, it contained varying ingredients of flattery, cajolery, and attempted justification; but it also was compounded by magic—the attempt not to persuade nature, but to coerce it. Though the authorities (*e.g., c.* 451–450 BC, Law of the Twelve Tables) sought to limit its noxious aspects, magic continued to abound throughout the ancient world. Even official rites remained full of its survivals, notably the annual festival of the Lupercalia and the ritual dances of the Salii in honour of Mars. Romans in historical times regarded magic as an oriental intrusion, but Italian tribes, such as the Marsi and Paeligni, were famous for such practices. Among them curses figured prominently, and curse inscriptions from *c.* 500 BC onward have been found in large numbers. There were also numerous survivals of taboo, a negative branch of magic: people were admonished to have no dealings with strangers, corpses, newborn children, spots struck by lightning, etc., lest harm would befall them.

Religion in the Etruscan period. The apparent amalgamation of the Latin and Sabine villages of Rome coincided with, or more probably was soon followed by, a period in which Rome was under the control of at least one dynasty (the Tarquins) from Etruria, north of the Tiber (*c.* 575–510 BC, though some scholars would extend this domination to *c.* 450).

Importance of ritual. The Etruscans felt profound religious anxieties and were more devoted to ritual than any other people of the ancient Western world. Though sources are, again, late and unsatisfactory, it appears that they possessed a comprehensive collection of rules regulating these rites. Etruscan culture was heavily based on influences from Greece in its orientalizing period, conveyed mainly through Greek centres (such as Cumae) in Campania, colonized by Euboeans, who were also prominent in Syrian markets. But the religion of Etruria proclaims a very un-Greek view of the abasement and nonentity of man before the gods and their will.

Divination and views of the afterlife

To the Etruscans the whole fanatical effort of life was directed toward forcing their deities, led by Tinia or Tin (Jupiter), to yield up their secrets by divination. They saw an intimate link existing between heaven and earth, which seemed to echo one another within a unitary system, and they were more ambitious than either Greeks or Romans in their claims to foretell the future. They also formed an exceptionally complex, rich, and imaginative picture of the afterlife. The living were perpetually obsessed by their care for the dead, expressed in elaborate, magnificently equipped and decorated tombs and lavish sacrifices. For, in spite of beliefs in an underworld, or Hades, there was also a conviction that the individuality of the dead somehow continued in their mortal remains; and it was therefore imperative that they take pleasure in their graves or tombs and not return to haunt the living. From the 4th century BC onward, after the Etruscans had lost their political power to Rome, their art depicts horrors indicating an increasing fear of what death might bring.

Influence on Roman religion. The Roman religion continued to display certain obvious debts to the period when the city had been under Etruscan control. It is true that the Roman shades (Di Manes) were much less substantial than the fantastic Etruscan conceptions and, although Etruscan divination by the liver and entrails survived and later became increasingly fashionable in Rome, Roman diviners in general, products of a more realistic and prosaic society, never aspired to such precise information about the future as the Etruscans had hoped to gain. Yet, it was the Etruscans who first gave a vigorous definition to Italian religious forms. Indeed, many of the religious features that patriotic historians preferred to ascribe to the mythical King Numa Pompilius (who was supposed to have been Romulus' Sabine successor in the 8th century BC—the man of peace following the man of war) date, in fact, from the period of Etruscan domination two centuries later. Nevertheless, Romans acknowledged a debt to Etruria that included much ceremony and ritual and the plan, appearance, and decoration of a number of temples, notably the great shrine of the Capitoline Triad, Jupiter, Juno, and Minerva. The Romans also were indebted to the Etruscans for their first statues of gods, including the cult image of Jupiter commissioned from an Etruscan for the Capitoline temple. Such statuary, showing the gods in human shape, encouraged the Romans to think of their gods in this way, with the consequent possibility of investing them with myths, which thereafter gradually accumulated around them in the form of Hellenic stories often infused with a native patriotic element.

The calendar

Above all, Rome owed to its Etruscan kings its religious calendar. In addition to poetical works discussing the calendar in antiquarian fashion, such as the *Fasti* of Ovid, there are extant fragments of about 40 copies of the calendar itself, in a revised shape established by Julius Caesar. Besides the Julian revision, there is an incomplete pre-Caesarian, Republican calendar, the Fasti Antiates, discovered at Antium (Anzio); it dates from after 100 BC. It is possible to detect in these calendars much that is very ancient, including a pre-Etruscan 10-month solar year. However, the basis of the calendars, in their surviving form, is later, since it consists of an attempt to reconcile the solar and lunar year, in accordance with Babylonian calculations. This endeavour belongs to the period of Etruscan domination of Rome—for example, the names of the months April and June (in their Roman form) come from Etruria. Moreover, the presence or absence of certain festivals permits a dating approximating to the time of Etruscan domination in the later 6th century BC. Additional modifications were introduced in the following century and again when the calendar was subsequently published (30 BC).

The festivals it records, of which the earliest are indicated in large letters, reflect a period of transition between country and town life. Though local cult continued to remain active, many forms of worship hitherto maintained by families and farms had now been taken over by the comparatively mature Roman state. The state management blocked any tendency toward spiritualization and removed the need for any vigorous individual participation; however, by ensuring that the gods were conciliated by a schedule corresponding to the regular process of nature, it made the individual citizens feel for centuries that relations with the supernatural were being maintained safely.

Religion in the early Republic. Even if, as tradition records, a coup d'état dislodged the Etruscan kings before 500 BC, in the first half of the 5th century there was no weakening of trade relations with Etruria. Its southern cities, such as Caere (Cerveteri) and Veii close to Rome, had long used the Greek city of Cumae as a commercial outlet, converting it into an important grain supplier. And now Rome, faced with a shortage of grain, arranged for it to be imported from Cumae. The same city also influenced the foundation of Roman temples in the Greek style. Rome, which had already become accustomed to Greek religious customs in the Etruscan epoch, now showed a willingness to absorb them. This forms a strange contrast to its deeply ingrained religious conservatism. Moreover, at some quite early stage (though there is no positive evidence of the practice until the 3rd century), Romans borrowed from elsewhere in Italy a special ritual (*evocatio*) for inviting the patron deities of captured towns to abandon their homes and migrate to Rome.

Influence of Greek religion

In an emergency in 399 BC, during a difficult siege of Veii, Rome carried Hellenization further by importing a Greek rite in which, as an appeal to emotional feeling, images of pairs of gods were exhibited on couches before tables spread with food and drink; this rite (*lectisternium*) was designed to make them Rome's welcome guests. From the same century onward, if not earlier, pestilences were averted by another ritual (*supplicatio*), in which the whole populace went around the temples and prostrated themselves in Greek fashion. Later the custom was extended to the celebration of victories.

Religion in the later Republic: crises and new trends. The *lectisternium* was repeated, with increased elaboration and pomp, in 217 BC during a period in which emotional religion was running rampant because of Hannibal's invasion of Italy in the Second Punic War. Faced with a flood of fears and anxieties and reports of many alarming and extraordinary events, Rome took precautions to secure the favour of all manner of gods. Among them, as a desperate attempt at novelty when appeals to the usual deities seemed stale, was the introduction of the Great Mother of Asia Minor, Cybele (204 BC). Eighteen years later, the equally orgiastic worship of Dionysus (Bacchus) was coming in so rapidly and violently, by way of southern Italy, that the Senate, scenting subversion, repressed its practitioners. But these and other mystery religions, promising initiation, afterlife, and an excitement that Roman national cults could not provide, had come to stay and, although there were long periods of official disapproval before acclimatization was completed, they gradually played an immense part upon the religious scene. Eastern astrology, too, became extremely popular. It was based on the conviction that, since there is cosmic sympathy between the earth and other heavenly bodies, and since, therefore, the emanations of these bodies influence the earth, men must learn how to foresee their dictates—and outwit them.

Influence of Stoicism

Astrological practices received encouragement from Stoic philosophy, which was introduced to Rome in the 2nd and early 1st centuries BC, notably by Panaetius and Poseidonius. The Stoics saw this pseudoscience as proof of the Platonic unity of the universe. Stoicism affected Roman religious thinking in at least three other ways. First, it had a deterministic effect, encouraging a widespread belief in Fate and also, somewhat illogically, in Fortune, both of which were revered in other parts of the Mediterranean and Middle Eastern world. Second, Stoicism infused a new spirituality into religious thinking by its insistence that the human soul is part of the universal spirit and shares its divinity. Third, the moral implication of this, as the Stoics pointed out, was that all men are brothers and must treat each other accordingly. This demonstration struck a chord in the psychology of the Romans, who possessed strongly ethical inclinations and now, at last, saw this trend supported and justified by a philosophical sanction that their formalistic religion had not provided. In changing times of imperialism, materialism, and widespread heart-searching, the state religion had failed to fill the vacuum, and philosophy stepped in instead. At the same time the negative approach of Roman religion to the afterlife was counteracted by an influx of speculations that blended theology, mysticism, and magic and claimed the mythical Orpheus and the part historical, part legendary Pythagoras as prophets.

While their national poet Ennius helped to diffuse such beliefs, he and the comic dramatist Plautus ridiculed the traditional Roman gods on the stage. The upper-class attitude of the times was expressed by the historian Polybius, the priestly lawyer Scaevola, the scholarly Varro, and the orator and philosopher Cicero, who maintained that the importance of religion was political, residing in its power to keep the multitude under control, to prevent social chaos, and to promote patriotic feeling.

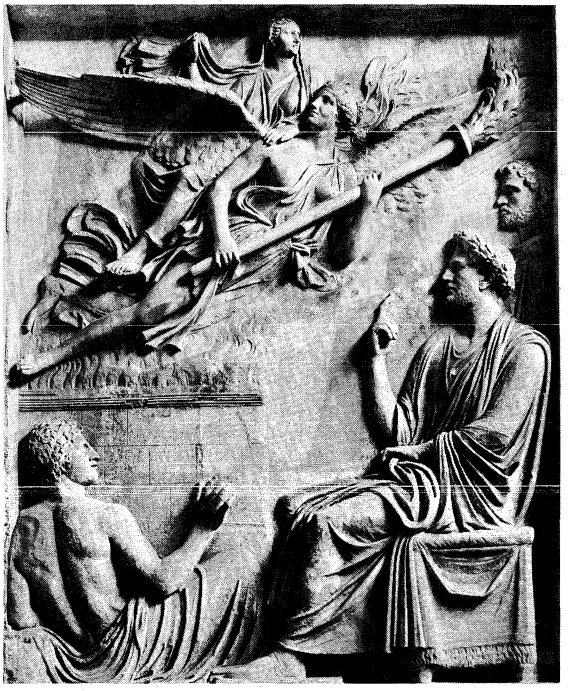
The imperial epoch: the final forms of Roman paganism. After the prolonged horrors of civil war had ended (30 BC), the victorious Octavian, the adoptive son of the dictator Caesar and founder of the imperial regime or principate, decided, correctly, that the ancient religion was far from dead and that the restoration of all its forms would respond to a strong popular, instinctive belief that the disasters of the past generations had been due to the neglect of religious duties.

Deification of Caesar and Augustus

The imperial cult. Octavian himself took the name Augustus, a term indicating a claim to reverence. This did not make him a god in his lifetime, but, combined with the insertion of his *numen* and his *genius* (originally the procreative power that enables a family to be carried on) into certain cults, it prepared the way for his posthumous deification, just as Caesar had been deified before him. Both were deified by the state because they seemed to have given Rome gifts worthy of a god. From earliest times in Greece there had been an idea that, if someone saved you, you should pay him the honours you would offer to a god. Alexander the Great and his successors had demanded reverence as divine saviours, and Ptolemy II Philadelphus of Egypt introduced a cult of his own living person. The Stoic belief that the human soul was part of the world soul was a corollary of the view that great men possessed a larger share of this divine element. Moreover, the 3rd-century-BC mythographer Euhemerus had elaborated a theory that the gods themselves had once been human; this idea was readily adapted to the supposed careers of Heracles (Hercules) and the Dioscuri (Castor and Polydeuces [Pollux]); and the Romans applied it to their own gods Saturn and Quirinus, the latter identified with the national founder, Romulus, risen to heaven. And so it became customary—if emperors (and empresses) were approved of in their lives—to raise them to divinity after their deaths. They were called *divi,* not *dei* like the Olympian gods; the latter were prayed to, but the former were regarded with veneration and gratitude.

As the empire proceeded and the old religion seemed more and more irrelevant to people's personal preoccupations and successive national emergencies, the cult of the *divi,* subsequently grouped together in a single Hall of Fame, remained foremost among the patriotic cults that were increasingly encouraged as unifying forces. Concentrating on the protectors of the emperor and the nation, they included the worship of Rome herself, and of the *genius* of the Roman people; for the army a number of special military celebrations are recorded on the Calendar of Doura-Europus in Mesopotamia (Feriale Duranum, *c.* AD 225–27). As for the ruling emperors, they were more and more frequently treated as divine, with varying degrees of formality, and officially they often were compared with gods. As monotheistic tendencies grew, however, this custom led not so much to their identification with the gods as to the doctrine that they were the elect of the divine powers, who were defined as their companions (*comites*). In pursuance of this way of thinking, as official paganism approached its last days, the emperors Diocletian and Maximian took the names Jovius and Herculius, respectively, after their Companions and Patrons Jupiter and Hercules.

Anderson—Alinari from Art Resources/EB Inc.



Apotheosis of Faustina, wife of Marcus Aurelius, ancient bas-relief. In the Capitoline Museum, Rome.

Introduction of Christianity and Mithraism. By now, however, the humanistic idea that men could become gods had ceased to have any plausibility. Plotinus and his Neoplatonism, the dominant philosophy of the pagan world from the mid-3rd century AD, had given powerful, mystical shape to the Platonic and Stoic conception that the universe is governed by a single force. On the other hand, the greatest religious figure of the century, the Iranian Mani, who had started to preach in Mesopotamia *c.* 240, dramatically preached the opposing dualistic idea that the world is the creation not only of a good power but of an evil one as well. Mani's church, which alarmed Diocletian and for a time attracted the great Christian theologian St. Augustine, absorbed many of the innumerable cults of Gnostics who claimed special knowledge (*gnōsis*) by illumination and revelation and taught how people can purge the nonspiritual from within themselves and escape their earthly prison. More impressively, the cult of the Persian Mithra blended the dualism of Mani with the emotional initiations of the mystery religions (corrected by a much sterner tone of moral endeavour) and became a strong link between the cult of the Sun (which appealed to contemporary monotheists) and the fashionable revulsion from the senses that was shortly to lead to Christian monasticism. Like Christianity, Mithraism had its sacraments; but the life of Mithra exercised a less far-reaching appeal than the life of Christ, and Mithra's cult excluded women.

Speculative religious thought

Christianity, unique in its universal charity and unique

also in its demand for a noble effort of faith in Jesus' blend of divinity and humanity, was the religion that prevailed in the Roman world. It satisfied the emperor Constantine's impulsive need for divine support, and from AD 312 onward, by a complex and gradual process, it became the official religion of the empire.

The survival of Roman religion. For a time, coins and other monuments continued to link Christian doctrines with the worship of the Sun, to which Constantine had been addicted previously. But even when this phase came to an end, Roman paganism continued to exert other, permanent influences, great and small. The emperors passed on to the popes the title of chief priest, *pontifex maximus.* The saints, with their distribution of functions, often seemed to perpetuate the many *numina* of ancient tradition. The ecclesiastical calendar retains numerous remnants of pre-Christian festivals—notably Christmas, which blends elements including both the feast of the Saturnalia and the birthday of Mithra. But, most of all, the mainstream of Western Christianity owed ancient Rome the firm discipline that gave it stability and shape, combining insistence on established forms with the possibility of recognizing that novelties need not be excluded, since they were implicit from the start.

BELIEFS, PRACTICES, AND INSTITUTIONS

The earliest divinities. The early Romans, like other Italians, worshiped not only purely functional and local forces but also certain high gods. Chief among them was the sky god Jupiter, whose cult, at first limited to the communities around the Alban Hills, later gained Rome as an adherent. The Romans gave Jupiter his own priest (*flamen*), and the fact that there were two other senior *flamines,* devoted to Mars and Quirinus, confirms other indications that the cults of these three deities, envisaged perhaps in some sort of association, belonged to a very early stratum (though the theory of their correspondence to the three-class social division of the early Indo-European-speaking peoples is generally unacceptable). Mars, whose name may or may not be Indo-European, was a high god of many Italian peoples, as liturgical bronze tablets found at Iguvium (Gubbio), the Tabulae Iguvinae (*c.* 200–*c.* 80 BC), confirm, protecting them in war and defending their agriculture and animals against disease. Later, he was identified with the Greek god of war, Ares, and also was regarded as the father of Romulus. Mars Gradivus presided over the beginning of a war and Mars Quirinus over its end, but earlier Quirinus had apparently, as a separate deity, been the patron of the Quirinal village before its amalgamation with the Palatine; subsequently he was believed to have been the god that Romulus became when he ascended into heaven.

Jupiter, Mars, and Quirinus

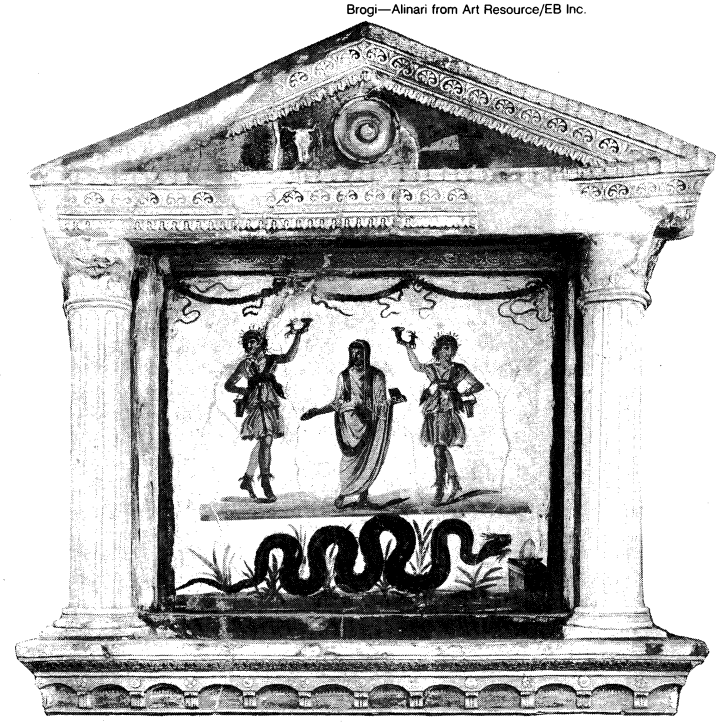
Janus and Vesta

Two other forces that belong to an early phase were Janus and Vesta, the powers of the door and hearth, respectively. Janus, who had no Greek equivalent, was worshiped beside the Forum in a small shrine with double doors at either end and originated either from a divine power that regulated the passage over running water or rather, perhaps, from sacred doorways like those found on the art of Bronze Age Mycenae. Janus originally stood for the magic of the door of a private house or hut and later became a part of the state religion. The gates of his temple were formally closed when the state was at peace, a custom going back to the primitive war magic that required armies to march out to battle by this properly sanctified route. Vesta, too, passed from the home to the state, always retaining a circular temple reminiscent of the primitive huts whose form can be reconstructed from traces left in the earth and from surviving funerary urns. Vesta's shrine contained the eternal fire, but the absence of a statue indicates that it preceded the anthropomorphic period; its correspondence with the Indian *garhapatya,* "house-father's fire," suggest an origin prior to the time of the differentiation of the Indo-European-speaking peoples. The cultic site just outside the area of the primitive Palatine settlement indicates that there had been a form of fire worship even earlier than Vesta's (dedicated to the deity Caca) on the Palatine itself. The cult of Vesta, tended by her Virgins, continued to flourish until the end of antiquity, endowed with an important role in the sacred protectorship of Rome.

Manes, Indigetes, Lares, and Penates

The Di Manes, collective powers (later "spirits") of the dead, may mean "the good people," an anxious euphemism like the Greek name of "the kindly ones" for the Furies. As a member of the family or clan, however, the dead man or woman would, more specifically, be one of the Di Parentes; reverence for ancestors was the core of Roman religious and social life. Di Indigetes was a name given collectively to these forebears, as well as to other deified powers or spirits who likewise controlled the destiny of Rome. For example, the name Indiges is applied to Aeneas, whose mythical immigration from Troy led to the eventual foundation of the city. According to an inscription of the 4th century BC (found at Tor Tignosa, 15 miles south of Rome), Aeneas is also called Lar, which indicates that the Lares, too, were originally regarded as divine ancestors and not as deities who presided over the farmland. The Lares were worshiped wherever properties adjoined, and inside every home their statuettes were placed in the domestic shrine (*lararium*). Under state control they moved from boundaries of properties to crossroads (where Augustus eventually associated his own *genius* with the cult) and were worshiped as the guardian spirits of the whole community (Lares Praestites). The cult of the Di Penates likewise moved from house to state. From very early times the Penates, the powers that ensured that there was enough to eat, were worshiped in every home. They also came to be regarded as national protectors, the Penates Publici. Originally they were synonymous with the Dioscuri. The legend that they had been brought to Italy by Aeneas with his followers from Troy was imported from Lavinium (Pratica di Mare) when the early Romans incorporated that town into their own state.

Brogi—Alinari from Art Resource/EB Inc.



Altar of the Lares, depicting two Lares on either side of the Genius, AD 69–72. In the House of the Vettii, Pompei.

The divinities of the later Regal period. Two other deities whose Roman cults tradition attributed to the period of the kings were Diana and Fors Fortuna. Diana, an Italian wood goddess worshiped at Aricia (Ariccia) in Latium and prayed to by women who wanted children, was in due course identified with the Greek Artemis. Her temple on the Aventine Hill (*c.* 540 BC) with its statue, an imitation of a Greek model from Massilia (Marseille), was based on the Temple of Artemis of Ephesus. By establishing such a sanctuary, the Roman monarch Servius Tullius hoped to emulate the Pan-Ionian League among the Latin peoples. Fors Fortuna, whose temple across the

Tiber from the city was one of the few that slaves could attend, was similar to the oracular shrines of Fortuna at Antium (Anzio) and Praeneste (Palestrina). Originally a farming deity, she eventually represented luck. She came to be identified with Tyche, the patroness of cities and goddess of Fortune among the Hellenistic Greeks.

In Roman tradition, Servius Tullius reigned between two Etruscan kings, Tarquinius Priscus and Tarquinius Superbus. The Etruscan kings began and perhaps finished the most important Roman temple, devoted to the cult of the Capitoline Triad, Jupiter, Juno, and Minerva (the dedication was believed to have taken place in 509 or 507 BC after the expulsion of the Etruscans). Such triads, housed in temples with three chambers (*cellae*), were an Etruscan institution. But the grouping of these three Roman deities seems to be owed to Greek anthropomorphic ideas, since Hera and Athena, with whom Juno and Minerva were identified, were respectively the wife and daughter of Zeus (Jupiter). In Italy, Juno (Uni in Etruscan) was sometimes the warlike high goddess of a town (*e.g.*, Lanuvium [Lanuvio] in Latium), but her chief function was to supervise the life of women, and particularly their sexual life. The functions of Minerva concerned craftsmen and reflected the growing industrial life of Rome. Two gods with Etruscan names, both worshiped at open altars before they had temples in Rome, were Vulcan and Saturn, the former a fire god identified with the Greek blacksmiths' deity Hephaestus, and the latter an agricultural god identified with Cronus, the father of Zeus. Saturn was worshiped in Greek fashion, with head uncovered.

Capitoline Triad

The focal point of the cult of Hercules was the Great Altar (Ara Maxima) in the cattle market, just inside the boundaries of the primitive Palatine settlement. The altar may be traced to a shrine of Melkart established by traders from Phoenicia in the 7th century BC. The name of the god, however, was derived from the Greek Heracles, whose worship spread northward from southern Italy, brought by traders who venerated his journeys, his labours, and his power to avert evil. In a market frequented by strangers, a widely recognized divinity of this type was needed to keep the peace. The Greek cult, at first private, perhaps dates from the 5th century BC.

The divinities of the Republic. An important series of temples was founded early in the 5th century BC. The completion of the temple of the Etruscan Saturn was attributed to this time (497). A shrine honouring the twin horsemen, the Dioscuri (Castor and Pollux), was also built in this period. An inscription from Lavinium describing them by the Greek term *kouroi* indicates a Greek origin (from southern Italy) without Etruscan mediation. In legend, the Dioscuri had helped Rome to victory in a battle against the Latins at Lake Regillus, and in historic times, on anniversaries of that engagement, they continued to preside over the annual parade of knights (*equites*). From southern Italy, too, came the cult of Ceres, whose temple traditionally was vowed in 496 and dedicated in 493. Ceres was an old Italian deity who presided over the generative powers of nature and came to be identified with Demeter, the Greek goddess of grain. She owed her installation in Rome to the influence of the Greek colony of Cumae, from which the Romans imported grain during a threatened famine. The association of Ceres at this temple with two other deities, Liber (a fertility god identified with Dionysus) and Libera (his female counterpart), was based on the triad at Eleusis in Greece. The Roman temple, built in the Etruscan style but with Greek ornamentation, stood beside a Greek trading centre on the Aventine Hill and became a rallying ground for the plebeians, the humbler section of the community who were hard hit by the grain shortage at this time and who were pressing for their rights against the patricians.

Ceres, Apollo, and Venus

Cumae also played a part in the introduction of Apollo. The Sibylline oracles housed in Apollo's shrine at Cumae allegedly were brought to Rome by the last Etruscan kings. The importation of the cult (431 BC) was prescribed by the Sibylline Books at a time when Rome, as on earlier occasions, had requested Cumae for help with grain. The Cumaean Apollo, however, was primarily prophetic, whereas the Roman cult, introduced at a time of epidemic, was concerned principally with his gifts as a healer. This role may possibly have been derived from the Etruscans, whose Apollo is known from a superb statue of *c.* 500 BC from Veii, Etruria's nearest city to Rome. In 82 BC the Sibylline Books were destroyed and replaced by a collection assembled from various sources. Later, Augustus elevated Apollo as the patron of himself and his regime, intending thereby to convert the brilliant Hellenic god of peace and civilization to the glory of Rome.

Unlike Apollo, Aphrodite did not keep her name when she became identified with an Italian deity. Instead, she took on the name Venus, derived, without complete certainty, from the idea of *venus,* "blooming nature" (the derivation from *venia,* "grace," seems less likely). She gained greatly in significance because of the legend that she was the mother of Aeneas, the ancestor of Rome, whom statuettes of the 5th century BC from Veii show escaping from Troy with his father and son. From the time of the Punic Wars 200 years later the Trojan legend grew, for long before the 1st-century-BC dictators Sulla and Caesar claimed Venus as their ancestor, the story was interpreted as the preface to the Carthaginian struggle.

A number of gods were spoken of as possessing accompaniments, often in the feminine gender; *e.g.,* Lua Saturni and Moles Martis. These attachments, sometimes spoken of as cult partners, were not the wives of the male divinities but rather expressed a special aspect of their power or will. A similar origin could be ascribed to the worship of divine powers representing "qualities." Fides ("Faith" or "Loyalty"), for example, may at first have been an attribute or aspect of a Latin-Sabine god of oaths, Semo Sanctus Dius Fidius; and in the same way Victoria may come from Jupiter Victor. Some of these concepts were worshiped very early, such as Ops ("Plenty," later associated with Saturn and equated with Hebe), and Juventas (who watched over the men of military age). The first of these qualities to receive a temple, as far as is known, is Concordia (367), in celebration of the end of civil strife. Salus (health or well-being) followed in *c.* 302, Victoria in *c.* 300, Pietas (dutifulness to family and gods, later exalted by Virgil as the whole basis of Roman religion) in 191. The Greeks, too, from the earliest days, had clothed such qualities in words; *e.g.,* Shame, Peace, Justice, and Fortune. In the Hellenic world they had a wide variety of signification, ranging from full-fledged divinity to nothing more than abstractions. But in early Rome and Italy they were in no sense abstractions or allegories and were likewise not thought of as possessing the anthropomorphic shape that the term personification might imply. They were things, objects of worship, like many other functions that were venerated. They were external divine forces working upon humans and affecting them with the qualities that their names described. Later on, under philosophical (particularly Stoic) influences that flooded into ethically minded Rome, they duly took their place as moral concepts, the Virtues and Blessings which abounded for centuries and were depicted in human form on Roman coinage as part of the imperial propaganda.

Divine qualities

The Sun and stars. Little or no contribution to cosmology was made in the Roman world, and the demonstration of Aristarchus of Samos (*c.* 270 BC) that the Earth revolves around the Sun received virtually no support. The complicated geocentric interpretation that held sway in Rome was summed up in Cicero's *Dream of Scipio.* It formed the basis for the concept of the solar system on which the popular pseudoscience of astrology was founded, the Sun being regarded as the centre of the concentric planetary spheres encircling the Earth—not the centre of the cosmos in the sense of Aristarchus but its heart. From the 5th century BC onward this solar god was identified with Apollo in his role as the supreme dispenser of agricultural wealth. Possessor of a sacred grove at Lavinium, Sol Indiges was regarded as one of the divine ancestors of Rome. During the last centuries before the Christian era, worship of the Sun spread throughout the Mediterranean world and formed the principal rallying point of paganism's last years. Closely associated with the sun cult was that of Mithra, the Sun's ally and agent who was elevated to partake of communion and the love feast as

Sol

the god's companion. Sun worship was popular in the army, and particularly on the Danube. Aurelian, one of the great military emperors produced by that area in the 3rd century, built a magnificent temple of Sol Invictus (the "Unconquered Sun") at Rome (274). Constantine the Great declared the Sun his Comrade on empire-wide coinages and devoted himself to the cult until he adopted Christianity in its stead.

The *rex sacrorum, flamines,* and colleges of priests

Priests. Precedence among Roman priests belonged to the *rex sacrorum* ("king of the sacred rites"), who, after the expulsion of the kings, took over the residue of their religious powers and duties that had not been assumed by the Republican officers of state. Nevertheless, the hold exercised by the *rex sacrorum* and his colleagues was weakened by the Law of the Twelve Tables (*c.* 451–450 BC), which displayed the secular arm exercising some control over sacral law. As late as *c.* 275 BC the religious calendar was still dated by the *rex sacrorum,* but by this time he was already fading into the background.

Very early origins can also be attributed to some of the *flamines,* the priests of certain specific cults, and particularly to the three major *flamines* of Jupiter, Mars, and Quirinus. Jupiter's priest, the *flamen dialis,* was encompassed by an extraordinary series of taboos, some dating to the Bronze Age, which made it difficult to fill the office in historic times.

Except for the *rex sacrorum* and *flamen dialis,* whose duties were unusually professional and technical, almost all Roman priesthoods were held by men prominent in public life. The social distinction and political prestige carried by these part-time posts caused them to be keenly fought for.

There were four chief colleges, or boards, of priests: the *pontifices, augures, quindecimviri sacris faciundis,* and *epulones.* Originally three, and finally 16 in number, the *pontifices* (whose name may recall antique tasks and magic rites in connection with bridges) had assumed control of the religious system by the 3rd century BC. The chief priest, the *pontifex maximus* (the head of the state clergy), was an elected official and not chosen from the existing *pontifices.* The *augures,* whose name may have been derived from the practice of magic in fertility rites and perhaps meant "increasers," had the task of discovering whether or not the gods approved of an action. This they performed mainly by interpreting divine signs in the movements of birds (*auspicia*). Such divination was elevated, perhaps under Etruscan influence, into an indispensable preliminary to state acts, though the responsibility for the decision rested not with the priests but with the presiding state officials, who were said to "possess the auspices." In private life too, even as late as Cicero and Horace in the 1st century BC, important courses of action were often preceded by consultation of the heavens. The Etruscan method of divining from the liver and entrails of animals (*haruspicina*) became popular in the Second Punic War, though its practitioners (who numbered 60 under the empire) never attained an official priesthood.

Of the other two major colleges, the *quindecimviri* ("Board of Fifteen," who earlier had been 10 in number) *sacris faciundis* looked after foreign rites, and the members of the other body, the *epulones,* supervised religious feasts. There were also *fetiales,* priestly officials who were concerned with various aspects of international relationships, such as treaties and declarations of war. Also six Vestal Virgins, chosen as young girls from the old patrician families, tended the shrine and fire of Vesta and lived in the House of Vestals nearby, amid a formidable array of prehistoric taboos.

Annual festivals

Shrines and temples. The Roman calendar, as introduced or modified in the period of the Etruscan kings, contained 58 regular festivals. These included 45 Feriae Publicae, celebrated on the same fixed day every year, as well as the Ides of each month, which were sacred to Jupiter, and the Kalends of March, which belonged to Mars. Famous examples of Feriae Publicae were the Lupercalia (February 15) and Saturnalia (December 17, later extended). There were also the Feriae Conceptivae, the dates of which were fixed each year by the proper authority, and which included the Feriae Latinae ("Latin Festival") celebrated in the Alban Hills, usually at the end of April.

Templum is a term derived from Etruscan divination. First of all, it meant an area of the sky defined by the priest for his collection and interpretation of the omens. Later, by a projection of this area onto the earth, it came to signify a piece of ground set aside and consecrated to the gods. At first such areas did not contain sacred buildings, but there often were altars on such sites, and later shrines. In Rome, temples have been identified from *c.* 575 BC onward, including not only the round shrine of Vesta but also a group in a sacred area (S. Omobono), close to the river Tiber beside the cattle market (Forum Boarium). The great Etruscan temples, made of wood with terra-cotta ornaments, were constructed later and culminated in the temple of the Capitoline Triad. Subsequently, more solid materials, such as tuff (tufa), travertine, marble, cement, and brick, gradually came into use. Temple archives, now vanished, play a large part in the historical tradition, and the anniversaries of the vows to build the temples and their dedication were scrupulously remembered and celebrated on numerous coins.

Sacrifice and burial rites. The characteristic offering of the Romans was a sacrifice accompanied by a prayer or vow. (The Triumph, associated with Jupiter, was regarded as a thanksgiving in discharge of a vow.) Animal sacrifices were regarded as more effective than anything else, the pig being the commonest victim, with sheep and ox added on important occasions. Considered best of all were the basic elements of life: heart, liver, and kidneys. Human sacrifice, on the whole, was extraneous to Roman custom, though its practice among the Etruscans may have contributed to the institution of gladiatorial funeral games in both Etruria and Rome, and it was resorted to in major crises, notably during the Second Punic War (216 BC). Earlier in the century, and perhaps once before, a member of the family of the Decii had given up his life by self-sacrifice (*devotio*) in a critical battle.

Although ancestors were meticulously revered, there was nothing resembling the comprehensive Etruscan attention to the dead. In spite of elaborate philosophizing by Cicero and Virgil about the possibility of some sort of survival of the soul (especially for the deserving), most Romans' ideas of the afterlife, unless they believed in the promises of the mystery religions, were vague. Such ideas often amounted to a cautious hope or fear that the spirit in some sense lived on, and this was sometimes combined with an anxiety that the ghosts of the dead, especially the young dead who bore the living a grudge, might return and cause harm. Graves and tombs were inviolable, protected by supernatural powers and by taboos. In the earliest days of Rome both cremation and inhumation were practiced simultaneously, but by the 2nd century BC the former had prevailed. Some 300 years later, however, there was a massive reversion to inhumation, probably because of an inarticulate revival of the feeling that the future welfare of the soul depended on comfortable repose of the body—a feeling that, as sarcophagi show, was fully shared by the adherents of the mystery cults, though, on the rational level, it contradicted their assurance of an afterlife in some spiritual sphere. The designs on these tombs reflect the soul's survival as a personal entity that has won its right to paradise.

Cremation and inhumation

Religious art. A vast gallery of architecture, sculpture, numismatics, painting, and mosaics illustrates Roman religion and helps to fill the gaps left by the fragmentary, though extensive, literary and epigraphic record. Starting with primitive statuettes and terra-cotta temple decorations, this array eventually included masterpieces such as the Apollo of Veii. Other works of art, more than 400 years later, include paintings illustrating Dionysiac mysteries at Boscoreale near Pompeii, and the reliefs of Augustus' Ara Pacis at Rome; and with the Christian emblems of Constantinian sarcophagi and coinage a thousand years of ancient Roman religious art comes to an end.

CONCLUSION

Though Roman religion never produced a comprehensive code of conduct, its early rituals of house and farm engen-

dered a feeling of duty and unity. Its idea of reciprocal understanding between man and god not only imparted the sense of security that Romans needed in order to achieve their successes but stimulated, by analogy, the concept of mutual obligations and binding agreements between one person and another. Except for rare aberrations, such as human sacrifice, Roman religion was unspoiled by orgiastic rites and savage practices. Moreover—unlike ancient philosophy—it was neither sectarian nor exclusive. It was a tolerant religion, and it would be difficult to think of any other whose adherents committed fewer crimes and atrocities in its name. (M.Gr.)

Hellenistic religions

The period of Hellenistic influence (which extended roughly from 300 BC to AD 300), when taken as a whole, constitutes one of the most creative periods in the history of religions. It was a time of spiritual revolution in the Greek and Roman empires, when old cults died or were fundamentally transformed and when new religious movements came into being.

NATURE AND SIGNIFICANCE

The historical Hellenistic Age is defined as the period from the death of the Greco-Macedonian conqueror Alexander the Great (323 BC) to the conquest of Egypt by Rome (30 BC), but the influence of the Hellenistic religions extended to the time of Constantine, the first Christian Roman emperor (d. AD 337); these religions are confined to those that were active within the Mediterranean world. The empire of Alexander and his successors created a great world community which, whether in Macedonian, Greco-Roman, or its later Christian form, established a cultural unity that was destined to be broken only 1,000 years later with the advent of Muslim imperialism (beginning in 7th century AD). This empire was so vast as truly to stagger the imagination. Extending from the Strait of Gibraltar to the Indus River, from the forests of Germany and the steppes of Russia to the Sahara Desert and the Indian Ocean, it took in an area of some 1.5 million square miles (3.9 million square kilometres; most of Europe, the Mediterranean, the Middle East, Africa, Persia, and the borderlands of India) and had a total population of more than 54 million.

The study of Hellenistic religions is a study of the dynamics of religious persistence and change in this vast and culturally varied area. Almost every religion in this period occurred in both its homeland and in diasporic centres—the foreign cities in which its adherents lived as minority groups. For example, Isis (Egypt), Baal (Syria), the Great Mother (Phrygia), Yahweh (Palestine), and Mithra (Kurdistan) were worshiped in their native lands as well as in Rome and other cosmopolitan centres. With few exceptions, each of these religions, originally tied to a specific geographic area and people, had traditions extending back centuries before the Hellenistic period. In their homeland they were inextricably tied to local loyalties and ambitions. Each persisted in its native land with little perceptible change save for its becoming linked to nationalistic or messianic movements (centring on a deliverer figure) seeking to overthrow Greco-Roman political and cultural domination. Indeed, many of these native religions underwent a conscious archaism during this period, attempting to recover earlier forms and practices. Old texts in native languages (especially those related to relevant themes such as kingship) were recopied, national temples were restored, and old, mythic traditions were revived. From Palestine to Persia one may trace the rise of Wisdom literature (the teachings of a sage concerning the hidden purposes of the deity) and apocalyptic traditions (referring to a belief in the dramatic intervention of a god in human and natural events) that represent these central concerns—*i.e.,* national destiny, the importance of traditional lore, the saving power of kingship, and the revival of mythic images. Each of these native traditions likewise underwent hellenization (modifications based on Greek cultural ideas), but in a manner frequently different from their diasporic counterparts.

The dispersion of local and area cults and the resultant changes

Each of these native religions also had diasporic centres that exhibited marked change during the Hellenistic period. There was a noticeable lessening of concern on the part of the members of the dispersed religious group for the destiny and fortunes of the native land and also a relative severing of the traditional ties between religion and the land. Certain cult centres remained sites of pilgrimage or objects of sentimental attachment; but the old beliefs in national deities and the inextricable relationship of the deity to certain sacred places was weakened. Rather than a god who dwelt in his temple, the diasporic traditions evolved complicated techniques for achieving visions, epiphanies (manifestations of a god), or heavenly journeys to a transcendent god. This led to a change from concern for a religion of national prosperity to one for individual salvation, from focus on a particular ethnic group to concern for every human. The prophet or saviour replaced the priest and king as the chief religious figure. In the diasporic centres, as is generally characteristic of immigrant groups, there were two circles. The first (or inner circle) was composed of devout, full-time adherents of the cult for whom the deity retained a separate and decisive identity (*e.g.,* those of Yahweh, Zeus Sarapis, and Isis). Its membership was drawn from the ethnic group for whom the deity was indigenous, and the group tended to continue to speak the native language. The second (or outer circle) was composed of either second- and third-generation immigrants or converts from groups for whom the religion was not native. These individuals tended to speak Greek, and this began the lengthy process of reinterpretation of the archaic religion. Ancient sacred books were translated or paraphrased into Greek—*e.g.,* the 4th–3rd-century-BC Babylonian priest Berosus' version of Babylonian materials, the 4th–3rd-century-BC Egyptian priest Manetho's Egyptian accounts, the Jewish Septuagint (Greek version of the Old Testament), or the 1st-century-AD Jewish historian Josephus' *Antiquities of the Jews,* and the ethnic histories of the 1st-century-BC Greek writer Alexander Polyhistor. In each case the material was reinterpreted both in light of common Hellenistic ideals and in accord with the special traditions and needs of the diasporic community. Both the inner and outer circles fostered esotericism (secrets to be known only by initiates)—the former by its use of native language and its oral recollection of traditions from the homeland; the latter by its use of allegory and other similar methods to radically reinterpret the sacred texts. The difference between these groups was responsible for many shifts in the character of the religion. Most notable was the shift from elements characteristic of native religion in its definition of religion (*e.g.,* local tradition and custom, informal knowledge orally transmitted, and birth) to formulated dogma, creeds, law codes, and rules for conversion and admission that were characteristic of diasporic religion. It was a shift from "birthright" to "convinced" religion.

The history of Hellenistic religions is rarely the history of genuinely new religions. Rather it is best understood as the study of archaic Mediterranean religions in their Hellenistic phase within both their native and diasporic settings. It is usually by concentrating on the diaspora that the Hellenistic character of a cult has been described.

HISTORY

Religion from the death of Alexander to the reformation of Augustus: 323–27 BC. The conquests of Alexander opened the way for religious interchange between East and West; the political structures left behind by Alexander and continued by his successors provided strong incentives for the hellenization of native religions. Characteristic of this first period of Hellenistic religious history were the following developments: (1) the introduction of Oriental cults into the West, especially those associated with female deities who were either worshiped in frenzied rites of self-mutilation (*e.g.,* the Phrygian Cybele, brought to Rome in 204 BC; the Syrian Atargatis; or the Cappadocian Ma-Bellona) or in adoring contemplation of their beneficence and gentle rites of divine rebirth (*e.g.,* the Egyptian Isis, whose cult was widespread in the Greco-Roman world by the middle of the 2nd century BC); (2) the hellenization

Religious interchange between East and West

of native cults (most famously that of the archaic Egyptian god Sarapis whose Greek form was promulgated by Ptolemy I, the founder of the Egyptian Ptolemaic dynasty in 305 BC); (3) the development of the ideology of divine kingship based on Oriental kingship traditions; and (4) the rise of nationalistic and messianic movements directed against internal and external hellenization; *e.g.*, the Maccabean rebellion led by Judas Maccabeus against Jewish hellenizing parties and the Syrian overlords in 167–165 BC, and the numerous Egyptian rebellions, especially that led by the Egyptian independence leader Harmakhis in Thebais in 207/6 BC.

Religion from the Augustan reformation to the death of Marcus Aurelius: 27 BC–AD 180. Oriental cults underwent their most significant expansion westward during this period. Particularly noticeable was the success of a variety of prophets, magicians, and healers—*e.g.*, John the Baptist, Jesus, Simon Magus, Apollonius of Tyana, Alexander the Paphlagonian, and the cult of the healer Asclepius—whose preaching corresponded to the activities of various Greek and Roman philosophic missionaries. A developing tension between these "new" Eastern religions and the archaic Greco-Roman traditions was expressed internally in the attempt by the emperor Augustus to revive traditional Roman religious practices. Attempts were made to expel foreigners or to suppress foreign worship—*e.g.*, the suppression of the Bacchic mysteries (salvation cults devoted to the god Dionysus, or Bacchus) in Rome in 186 BC, or the numerous attempts to prohibit the worship of the Egyptian goddess Isis in Rome, beginning in 59 BC. The Augustan reformation also restored Roman sacred books and Greek temples.

Developing tensions between the Greco-Roman and the Eastern religions

Externally, the developing tension was expressed in wars, riots, and persecutions, such as the Jewish–pagan riots in Alexandria in AD 38 and 115–116, the Jewish–Roman wars of AD 66–70 and 132–135, and the beginning of the persecution of Christians under the Roman emperor Nero in AD 64. Another cause of tension was the elaboration of a full-blown cult of "emperor worship," beginning with the deification of Augustus (Sept. 17, AD 14) shortly after his death.

Religion from Commodus to Theodosius I: AD 180–395. After the death of the "philosopher-king" Marcus Aurelius in AD 180, his son Commodus became emperor, and a period of political instability began. The dominant feature of the concluding period of Hellenistic influence—and shortly thereafter—was the rapid growth of Christianity throughout the Roman Empire, culminating in the conversion to Christianity of the emperor Constantine in 313 and the religious legislation of the emperor Theodosius affirming in 380 the dogmas of the Christian Council of Nicaea—which had been convened in 325 under the auspices of Constantine—and prohibiting paganism in a decree of 392. In this period the various Hellenistic cults were victims of active hostilities, which were expressed through prohibition, acts of violence, and theological polemics between "pagans" and Christians (*e.g.*, the pagan philosophers Maximus of Tyre and Celsus, and the Christian philosophical theologians Irenaeus, Tertullian, and St. Clement of Alexandria, all of the 2nd century); but there were also brief periods of Hellenistic revitalization. The Neoplatonic school (based on a complicated system of levels of reality) of the 3rd-century philosophers Plotinus and Porphyry represented the culmination of Hellenistic religious philosophy. The Syrian solar cults of Sol Invictus (the "Unconquered Sun") and Jupiter Dolichenus played an important role under the emperors Antoninus Pius, the Severans—Septimius, and Alexander—and Elagabalus and these were hailed as the supreme deities of Rome under Aurelian, whose Sun temple was dedicated in 274. From Parthia, the dualistic and spiritual teachings of the 2nd-century Iranian prophet Mani were widely disseminated throughout the Empire. The Persian cult of the ancient Iranian god of light, Mithra, spread rapidly throughout the western and northern Empire during the 3rd through 5th centuries. Although these various traditions enjoyed brief imperial patronage under Julian, they eventually were subsumed under the political and religious hegemony of Christianity (see below).

BELIEFS, PRACTICES, AND INSTITUTIONS

The archaic religions of the Mediterranean world were primarily religions of etiquette. At the centre of these religions were complex systems governing the interrelationships between gods and humans, individuals and the state, and living people and their ancestors. The entire cosmos was conceived as a vast network of relationships, each component of which, whether divine or human, must know its place and fulfill its appointed role. The model for this all-encompassing system was the divine society of the gods, and the map of this system was the order of the planets and stars. Through astrology, divination, and oracles, people discerned the unalterable patterns of destiny and sought to bring their world (the microcosm) into harmony with the divine cosmos (the macrocosm; see also OCCULTISM: *Divination: astrology*).

The centrality of the relationship between the gods and men

Cincinnati Art Museum, Museum Purchase

Limestone relief of the goddess Tyche in a zodiac, 2nd century AD. In the Cincinnati Art Museum.

This archaic pattern of affirming and celebrating the order of the cosmos was expressed in the typical creation myth of the Middle Eastern and Mediterranean world, which consisted of a creation by combat between the forces of order and chaos. Order was understood to be something won in the beginning by the gods, and it was this primordial act of salvation that was renewed and reexperienced in the cult.

In the Hellenistic period a new religious world was experienced that required new religious expressions. The old religions of conformity and place no longer spoke to this new religious situation and its questions. What if the law and order of the cosmos was no longer seen as the creative expression of limits and the delineation of roles, but rather as an evil, perverse, confining structure from which man and the cosmos must escape? Rather than the archaic structures of celebration and conformity to place, the new religious mood spoke of escape and liberation from place and of salvation from an evil, imprisoned world. The characteristic religion of the Hellenistic period was dualistic. People sought to escape from the despotism of this world and its rulers (exemplified by the seven planetary spheres) and to ascend to another world of freedom. Hellenistic people saw themselves as exiles from their true home, the Beyond, and they sought for ways to return. They strove to regain their place in the world beyond this world where they truly belonged, to encounter the god beyond the god of this world who was the true god, and to awaken that part of themselves (their souls or spirits) that had descended from the heavenly realm by stripping off their bodies, which belonged to this world. The questions that the religions of the Hellenistic period sought to answer may be seen in a fragment from the 2nd-century Anatolian Gnostic teacher Theodotus: "What liberates is the

Salvation as liberation through knowledge of one's origin, identity, and destiny

knowledge of who we were [before our earthly existence] and what we have become [on earth]; where we were [the Beyond] and the place to which we have been thrown [the world]; where we are going and from what are we redeemed; what is birth and what is rebirth" (preserved in Clement of Alexandria, *Excerpta ex Theodoto,* 78.2).

The gods. In the Greco-Roman world during the Hellenistic period, archaic deities were transformed in part because of the new spirit of the age and in part by foreign influences. A number of the old chthonic (underworld) and agricultural (fertility) gods and the old agricultural mysteries (corporate renewal religions related to fertility concepts) fundamentally altered their character. Rather than an expression of the alternation of life and death, of fertility and sterility, and a celebration of the promise of renewal for the land and the people, the seasonal drama was homologized to a soteriology (salvation concept) concerning the destiny, fortune, and salvation of the individual after death. The collective agricultural rite became a mystery, a salvific experience reserved for the elect (such as the Greek mystery religion of Eleusis). Other traditions even more radically reinterpreted the ancient figures. The cosmic or seasonal drama was interiorized to refer to the divine soul within man that must be liberated. Such cults were dualistic mysteries distinguishing sharply between the body and soul. They taught that it is the soul alone that was initiated by passing through death or the Underworld, or by being dismembered so that it might be freed from the body and regain its rightful mode of spiritual existence (such as the Orphic—mystical—reinterpretation of the role of the agricultural god Dionysus). In the gnostic mysteries (the esoteric dualistic cults that viewed matter as evil and the spirit as good), this process was carried further through the identification of the experiences of the soul that was to be saved with the vicissitudes of a divine but fallen soul, which had to be redeemed by cultic activity and divine intervention. This view is illustrated in the concept of the paradoxical figure of the saved saviour, *salvator salvandus.*

Other deities, who had previously been associated with national destiny (*e.g.,* Zeus, Yahweh, and Isis), were raised to the status of transcendent, supreme deities whose power and ontological status (relating to being or existence) far surpassed the other gods, who were understood as their servants or antagonists. The religious person sought to make contact with, or to stand before, this one, true god of the Beyond. The piety of the individual was directed either toward preparing himself to ascend up through the planetary spheres to the realm of the transcendent god or toward calling the transcendent god down that he might appear to him in an epiphany or vision. These techniques for achieving ascent or a divine epiphany make up the bulk of the material that has usually been termed magical, theurgic (referring to the art of persuading a god to reveal himself and grant salvation, healing, and other requests), or astrological and that represents the characteristic expression of Hellenistic religiosity.

Individual piety

Cosmogony and cosmology. The cosmogonies (dealing with the origins of the world) and cosmologies (dealing with the ordering of the world) of the Hellenistic period centred around the problem of accounting for the distance between this world and the Beyond, or on accounting for the evil nature of this world and its gods. Many mythic schema were employed regarding the origin and ordering of this world. It was viewed as being: the result of the conscious or unconscious emanation from the transcendent realm; the result of the fall of a deity from the Beyond; the creation of a hostile, ignorant, or evil deity; or a joke or mistake. The purpose of this speculation was both pragmatic and soteriological: if one could determine how this creation came into being, one could reverse it or overcome it and be saved.

Religious organization. The temples and cult institutions of the various Hellenistic religions were repositories of the knowledge and techniques necessary for salvation and were the agents of the public worship of a particular deity. In addition, they served an important sociological role. In the new, cosmopolitan ideology that followed Alexander's conquests, the old nationalistic and ethnic boundaries had broken down and the problem of religious and social identity had become acute. The Hellenistic Age was characterized by the rapid growth of private religious societies (*thiasoi*). Though some were organized according to national origin or trade, the majority were dedicated to the worship of a particular deity. In many instances these groups began as immigrant associations (*e.g.,* an Egyptian association of devotees of Amon was chartered in Athens at the beginning of the 3rd century BC); but they often transcended these origins and became a new form of religious organization in which citizens of various countries, freemen and slaves, could be united by their common devotion and share in a common religious heritage. Admission to such groups was voluntary (in contradistinction to the archaic national or familial religious organizations) and demanded the payment of dues, submission to collective authority, and the acceptance of strict codes of morality. Most of these groups had regular meetings for a communal meal that served the dual role of sacramental participation (referring to the use of material elements believed to convey spiritual benefits among the members and with their deity) and the social function of fellowship; *i.e.,* the security of membership in a group and a shared sense of identity.

Hellenistic religions as voluntary associations

THE INFLUENCE OF HELLENISTIC RELIGIONS

The archaic gods worshiped during the Hellenistic period possessed a remarkable longevity. The Eleusinian Mysteries, founded in the 15th century BC, ceased in the 4th century AD; Dionysus, whose name first appears on tablets dated to *c.* 1400 BC, was last celebrated in the beginning of the 6th century AD; the last temple of Isis, whose cult extended back to the 2nd millennium BC in Egypt, was closed in AD 560. Yet even after these ceased as objects of devotion in the post-Constantinian period, they continued to exercise their influence. Hellenistic philosophy (Stoicism, Cynicism, Neo-Aristotelianism, Neo-Pythagoreanism, and Neoplatonism) provided key formulations for Jewish, Christian, and Muslim philosophy, theology, and mysticism through the 18th century. Hellenistic magic, theurgy, astrology, and alchemy remained influential until modern times in both East and West. Theosophy and other forms of the occult, especially since the Renaissance, drew their inspiration from the Hellenistic mystery cults, Hermeticism (Greco-Egyptian astrological, magical, and occultic movement), and Gnosticism. Various Jewish, Christian, and Muslim sectarian groups continued the theologies of many of the Hellenistic religions (especially dualistic modes of thought). Hellenistic sacred art and architecture has remained a basis of Christian and Jewish iconography and architecture to the present day. Figures such as Alexander the Great inspired a vast body of religious literature, especially in the Middle Ages. Many of the symbols and legends associated with Hellenistic deities persisted in folk literature and hagiography (stories of saints and "holy" persons). The basic forms of worship of both the Jewish and Christian communities were heavily influenced in their formative period by Hellenistic practices, and this remains fundamentally unchanged to the present time. Finally, the central religious literature of both traditions—the Jewish Talmud (an authoritative compendium of law, lore, and interpretation), the New Testament, and the later patristic literature of the early Church Fathers—are characteristic Hellenistic documents both in form and content. (J.Z.S.)

BIBLIOGRAPHY

General: Scholarly articles in English on the topics discussed below, with bibliographies, may be found in *The Encyclopedia of Religion,* ed. by MIRCEA ELIADE, 14 vol. (1987). The classic work on old European religion is MARIJA GIMBUTAS, *The Goddesses and Gods of Old Europe, 6500–3500 BC: Myths and Cult Images,* new and updated ed. (1982). Her theory is also condensed in two of her articles in *The Encyclopedia of Religion*: "Prehistoric Religions: Old Europe," vol. 11, pp. 506–515, and "Megalithic Religion: Prehistoric Evidence," vol. 9, pp. 336–344. For Gimbutas' updated bio-bibliography, see SUSAN NACEV SKOMAL and EDGAR C. POLOMÉ (eds.), *Proto-Indo-European: The Archaeology of a Linguistic Problem* (1987). The problem of the Indo-European homeland is discussed in, among others, PEDRO BOSCH GIMPERA, *El problema indoeuropeo*

(1960); GIACOMO DEVOTO, *Origini indeuropee* (1962); VLADIMIR I. GEORGIEV, *Introduction to the History of the Indo-European Languages,* trans. from Bulgarian (1981); and EDGAR C. POLOMÉ (ed.), *The Indo-Europeans in the Fourth and Third Millennia* (1982). The theories of the scholars T.V. Gamkrelidze and V.V. Ivanov, together with critical reactions from I.M. Diakonov and Marija Gimbutas, were made available by EDGAR C. POLOMÉ, "Recent Russian Papers on the Indo-European Problem and on the Ethnogenesis and Original Homeland of the Slavs," a special issue of *The Journal of Indo-European Studies,* 13(1–2) (Spring-Summer 1985).

A fine survey of scholarly theories on Indo-European religion is given by C. SCOTT LITTLETON, "Indo-European Religions: History of Study," in *The Encyclopedia of Religion,* vol. 7, pp. 204–213, but one should also keep in mind the nationalistic and racist theories connected with the names "Aryans" and "Indo-Germans," as discussed in LÉON POLIAKOV, *The Aryan Myth: A History of Racist and Nationalistic Ideas in Europe* (1974; originally published in French, 1971). The most comprehensive statement of the "tripartite ideology" is GEORGES DUMÉZIL, *L'idéologie tripartie des Indo-Européens* (1958). An analysis of Dumézil's work and theories can be found in C. SCOTT LITTLETON, *The New Comparative Mythology,* 3rd ed. (1982). On the "Dumézilian school," see especially two collections of essays: EDGAR C. POLOMÉ (ed.), *Homage to Georges Dumézil* (1982); and FRANÇOISE DESBORDES *et al., Pour un temps: Georges Dumézil* (1981), with contributions by many French and American scholars. (I.P.C.)

Celtic religion: JOHN RHYS, *Lectures on the Origin and Growth of Religion as Illustrated by Celtic Heathendom,* 3rd ed. (1898, reprinted 1979), although the classic work in English, is now out-of-date. Useful accounts include JOSEPH VENDRYÈS, ERNEST TONNELAT, and B.-O. UNBEGAUN, *Les Religions des Celtes, des Germains et des anciens Slaves* (1948); and PAUL-MARIE DUVAL, *Les Dieux de la Gaule,* new ed. updated and enlarged (1976). JOHN MACNEILL, *Celtic Religion* (1911?), provides a brief outline for an overview of the subject. THOMAS F. O'RAHILLY, *Early Irish History and Mythology* (1946, reissued 1971), contains massive learning based on a great wealth of material, including some fanciful conclusions. MARIE-LOUISE SJOESTEDT, *Gods and Heroes of the Celts* (1949, reissued 1982; originally published in French, 1940), is an extremely perceptive reading of the heroic function in Celtic mythological tradition. JAN DE VRIES, *Keltische Religion* (1961), is a comprehensive survey, useful as a reference work. PROINSIAS MAC CANA, *Celtic Mythology* (1970), contains a concise presentation and evaluation of the evidence, with copious illustrations. CLAUDE STERCKS, *Éléments de cosmogonie celtique* (1986), contains a fine interpretive essay on the goddess Epona and related deities. (P.Mac.C.)

Germanic religion: JACOB GRIMM, *Teutonic Mythology,* 4 vol. (1883–88, reprinted 1976; originally published in German, 4th ed., 3 vol., 1875–78), is still a most valuable source. JAN DE VRIES, *Altgermanische Religionsgeschichte,* 2nd ed., 2 vol. (1956–57, reprinted 1970), is a thorough account of Germanic heathendom in Scandinavia, Germany, and England. GEORGES DUMÉZIL, *Gods of the Ancient Northmen* (1973; originally published in French, 1959), offers a short account of German mythology based on the author's view of the Indo-European heritage in Germanic religion. R.L.M. DEROLEZ, *De godsdienst der Germanen* (1959), surveys the gods and myths, with special attention to runic inscriptions; there is also a French translation, *Les Dieux et la religion des Germains* (1962), and a German translation, *Götter und Mythen der Germanen* (1963, reissued 1976). GABRIEL TURVILLE-PETRE, *Myth and Religion of the North: The Religion of Ancient Scandinavia* (1964, reprinted 1975), gives a comprehensive account of Norse myth and religious practice. A.V. STRÖM and HARALDS BIEZAIS, *Germanische und baltische Religion* (1975), encompasses the whole development from prehistoric times to the conversion to Christianity, with somewhat controversial interpretations. RÉGIS BOYER, *La Religion des anciens Scandinaves: Yggdrasill* (1981), an original survey, covers the topic from the Bronze Age petroglyphs to the saga religion but is somewhat marred by inaccuracies. RUDOLF SIMEK, *Lexikon der germanischen Mythologie* (1984), is well documented and contains reliable information. JOHN LINDOW, *Scandinavian Mythology: An Annotated Bibliography* (1988), is excellent.

ROBERT J. GLENDINNING and HARALDUR BESSASON (eds.), *Edda: A Collection of Essays* (1983), provides valuable insight. The best English version remains LEE M. HOLLANDER (trans.), *The Poetic Edda,* 2nd ed. rev. (1962, reprinted 1986). For Snorri's presentation of Scandinavian mythology, the major source is SNORRI STURLUSON, *Gylfaginning,* ed. by GOTTFRIED LORENZ (1984), with a substantial commentary in German. The best edition of the *Germania* by CORNELIUS TACITUS is the annotated German translation by ALLAN A. LUND (1988); for an English edition, see the translation by M. HUTTON (1970) in the *Loeb Classical Library, Latin Authors* series. An essay on early Germanic religion in the context of ancient Germanic culture can be found in EDGAR C. POLOMÉ, "Germantum und religiose Vorstellungen," in HEINRICH BECK (ed.), *Germanenprobleme in heutiger Sicht* (1986), pp. 267–297. (E.C.Po.)

Finno-Ugric religion: A comprehensive presentation can be found in LOUIS HERBERT GRAY, GEORGE FOOT MORE, and J.A. MACCULLOCH (eds.), *The Mythology of All Races,* vol. 4, *Finno-Ugric, Siberian,* by UNO HOLMBERG (1927, reprinted 1964). More recent surveys with extensive bibliographies include IVAR PAULSON, "Die Religionen der finnischen Völker," in IVAR PAULSON, ÅKE HULTKRANTZ, and KARL JETTMAR (eds.), *Die Religionen Nordeurasiens und der amerikanischen Arktis* (1962), pp. 145–303; and LAURI HONKO, "Religionen der finnisch-ugrischen Völker," in JES PETER ASMUSSEN, JØRGEN LAESSØE, and CARSTEN COLPE (eds.), *Handbuch der Religionsgeschichte,* vol. 1, trans. from Danish (1971), pp. 173–224. (L.O.H.)

Baltic religion: MARIJA GIMBUTAS, *The Balts* (1963), pp. 179–204, gives a concise summary. HANS BERTULEIT, "Das Religionswesen der alten Preussen mit litauisch-lettischen Parallelen," *Prussia,* vol. 25 (1924), is still the only complete review of the Old Prussian religion. A critical examination of sources and research may be found in HARALDS BIEZAIS, *Die Religionsquellen der baltischen Völker und die Ergebnisse der bisherigen Forschungen* (1954). For a comprehensive collection of historic records of the Prussian, Lithuanian, and Lettish religion, see WILHELM MANNHARDT, *Letto-preussische Götterlehre* (1936, reissued 1971). HARALDS BIEZAIS, *Die Hauptgöttinnen der alten Letten* (1955), *Die Gottesgestalt der lettischen Volksreligion* (1961), *Die himmlische Götterfamilie der alten Letten* (1972), *Lichtgott der alten Letten* (1976), and *Die baltische Ikonographie* (1985), are devoted to central problems of Baltic religion, with exhaustive bibliographies. (H.Bi.)

Slavic religion: Slavic mythology is outlined by JAN MÁCHAL, "Slavic," in LOUIS HERBERT GRAY, GEORGE FOOT MOORE, and J.A. MACCULLOCH (eds.), *The Mythology of All Races,* vol. 3 (1918, reissued 1964); and MYROSLAVA T. ZNAYENKO, *The Gods of the Ancient Slavs: Tatishchev and the Beginnings of Slavic Mythology* (1980). ALEKSANDER BRÜCKNER, *Mitologia słowiańska i polska,* 2nd ed. (1985), represents an attempt to furnish an Indo-European interpretation of Slavic paganism; for the critical side of the problem this work remains indispensable. For the archaeological aspects, see KARL SCHUCHHARDT, *Arkona, Rethra, Vineta: Ortsuntersuchungen und ausgrabungen,* 2nd rev. and enlarged ed. (1926). A descriptive exposition can be found in B.-O. UNBEGAUN, *Les Religions des Celtes, des Germains et des Slaves* (1948). An attempt to find in the folklore traces of a more ancient mythology was made by W.R.S. RALSTON, *The Songs of the Russian People, as Illustrative of Slavonic Mythology and Russian Social Life,* 2nd ed. (1872, reprinted 1970). A collection of materials and provocative suggestions in the same field is found in V.J. MANSIKKA, *Die Religion der Ostslaven* (1922, reissued 1967). For ethnography, see DMITRIJ ZELENIN, *Russische (ostslavische) Volkskunde* (1927); EDMUND SCHNEEWEIS, *Grundriss des Volksglaubens und Volksbrauchs der Serbokroaten* (1935); and PIERRE BOGATYREV, *Actes magiques, rites et croyances en Russie subcarpathique* (1929).

(E.G./Ed.)

Greek religion: General works include MARTIN P. NILSSON, *Greek Popular Religion* (1940, reissued as *Greek Folk Religion,* 1972), a sound and detailed survey, *Greek Piety* (1948, reissued 1969; originally published in Swedish, 1946), a general survey, *The Minoan-Mycenaean Religion and Its Survival in Greek Religion,* 2nd rev. ed. (1950, reprinted 1971), the best account of origins, and *Geschichte der griechischen Religion* (1941–50), the standard history; H.J. ROSE, *Ancient Greek Religion* (1928, reissued 1948), a brief but masterly sketch; W.K.C. GUTHRIE, *The Greeks and Their Gods* (1950, reprinted 1985), the best general account, and *The Religion and Mythology of the Greeks* (1961), a brief sound sketch of origins; and JOHN POLLARD, *Seers, Shrines, and Sirens: The Greek Religious Revolution in the Sixth Century B.C.* (1965). JANE ELLEN HARRISON, *Themis: A Study of the Social Origins of Greek Religion,* 2nd ed. (1927, reissued 1974), *Prolegomena to the Study of Greek Religion,* 3rd ed. (1922, reprinted 1973), and a sequel, *Epilegomena to the Study of Greek Religion* (1921, reissued 1962); and GILBERT MURRAY, *Five Stages of Greek Religion* (1925), are dependent on an anthropology that has gone out of favour, but much may still be learned from them and much has been borrowed from them without acknowledgement. WALTER BURKERT, *Homo Necans: The Anthropology of Ancient Greek Sacrificial Ritual and Myth* (1983; originally published in German, 1972), and *Greek Religion* (1985; originally published in German, 1977), have broken much new ground in discussing the origins of Greek religion. A.W.H. ADKINS, *Merit and Responsibility: A Study in Greek Values* (1960, reprinted 1975), and *Moral Values and Political Behaviour in Ancient Greece: From Homer to the End of the Fifth Century* (1972), include studies of the religious vocabulary of the Greeks.

The copious works of the "Paris school" together constitute an account of Greek religion that combines the structuralism of the French social anthropologist Claude Lévi-Strauss with a detailed attention to the phenomena furnished by the evidence of Greek religion, literature, philosophy, and art. A few examples include JEAN-PIERRE VERNANT, *Myth and Thought Among the Greeks* (1983; originally published in French, 1965), and *Myth and Society in Ancient Greece* (1980; originally published in French, 1974); MARCEL DETIENNE, *The Gardens of Adonis: Spices in Greek Mythology* (1977; originally published in French, 1972), and *Dionysus at Large* (1989; originally published in French, 1986); MARCEL DETIENNE and JEAN-PIERRE VERNANT, *Cunning Intelligence in Greek Culture and Society* (1978; originally published in French, 1974); PIERRE VIDAL-NAQUET, *The Black Hunter: Forms of Thought and Forms of Society in the Greek World* (1986; originally published in French, 1981); and JEAN-PIERRE VERNANT and PIERRE VIDAL-NAQUET, *Myth and Tragedy in Ancient Greece* (1988; originally published in French, 2 vol., 1972–86).

Works on oracles and divination include the authoritative W.R. HALLIDAY, *Greek Divination: A Study of Its Methods and Principles* (1913, reissued 1967); PIERRE AMANDRY, *La Mantique apollinienne à Delphes: essai sur le fonctionnement de l'oracle* (1950, reprinted 1975); H.W. PARKE and D.E.W. WORMELL, *The Delphic Oracle,* 2 vol. (1956); ROBERT FLACELIÈRE, *Greek Oracles,* 2nd ed. (1976; originally published in French, 1961); and H.W. PARKE, *Greek Oracles* (1967), and *The Oracles of Zeus: Dodona, Olympia, Ammon* (1967). Mysteries and eschatology are treated in ERWIN ROHDE, *Psyche: The Cult of Souls and Belief in Immortality Among the Greeks* (1925, reprinted 1987; originally published in German, 8th ed., 2 vol., 1921), the fundamental work; W.K.C. GUTHRIE, *Orpheus and Greek Religion: A Study of the Orphic Movement,* 2nd rev. ed. (1952, reissued 1967), the best work on Orphism; IVAN M. LINFORTH, *The Arts of Orpheus* (1941, reprinted 1973), a hypercritical account; E.R. DODDS, *The Greeks and the Irrational* (1951, reissued 1973), the best account since Rohde; GEORGE E. MYLONAS, *Eleusis and the Eleusinian Mysteries* (1961, reissued 1974), a good general survey; C. KERÉNYI, *Eleusis: Archetypal Image of Mother and Daughter* (1967, reprinted 1977), a psychological account; and W.F. JACKSON KNIGHT, *Elysion: On Ancient Greek and Roman Beliefs Concerning a Life After Death* (1970). Works on cults and festivals include LEWIS RICHARD FARNELL, *The Cults of the Greek States,* 5 vol. (1896–1909, reissued 1969), the best critical survey in English, and *Greek Hero Cults and Ideas of Immortality* (1921, reprinted 1970), a formal and critical account; MARTIN P. NILSSON, *Griechische Feste von religiösen Bedeutung* (1906, reprinted 1975), the standard work on non-Attic festivals; ARTHUR BERNARD COOK, *Zeus: A Study in Ancient Religion,* 3 vol. (1914–40, vol. 1–2 reprinted in 3 vol., 1964–65), a monumental compendium of all the evidence; LUDWIG DEUBNER, *Attische Feste* (1932, reissued 1969), the standard work on Attic festivals; EMMA J. EDELSTEIN and LUDWIG EDELSTEIN, *Asclepius: A Collection and Interpretation of the Testimonies,* 2 vol. (1945–46, reprinted in 1 vol., 1988), the best account in English; C. KERÉNYI, *Asklepios: Archetypal Image of the Physician's Existence* (1959; originally published in German, 1956), a psychological account; and LUDWIG DREES, *Olympia: Gods, Artists, and Athletes* (1968; originally published in German, 1967), a full, popular account of the festival. The art and architecture of Greek religion are treated in VINCENT SCULLY, *The Earth, the Temple, and the Gods: Greek Sacred Architecture,* rev. ed. (1979), a full if somewhat fanciful account of temple siting; HELMUT BERVE and GOTTFRIED GRUBEN, *Greek Temples, Theatres, and Shrines* (1963), a detailed survey of the chief buildings; and BIRGITTA BERGQUIST, *The Archaic Greek Temenos: A Study of Structure and Function* (1967), a scholarly survey.

W.H. ROSCHER, *Ausführliches Lexikon der griechischen und römischen Mythologie,* 6 vol. in 9 (1884–1937, reprinted 7 vol. in 10, 1977–78), is the authoritative encyclopaedia of Greek mythology. Other works on the subject include MARTIN P. NILSSON, *The Mycenean Origin of Greek Mythology* (1932, reissued 1983), a pioneer work, and *Cults, Myths, Oracles, and Politics in Ancient Greece* (1951, reprinted 1986), an excellent survey; C. KERÉNYI, *The Gods of the Greeks* (1951, reissued 1982; originally published in German, 1951), containing detailed data, and *The Heroes of the Greeks* (1959, reissued 1981; originally published in German, 1958), a dictionary of saga; H.J. ROSE, *A Handbook of Greek Mythology, Including Its Extension to Rome,* 6th ed. (1958, reissued 1972), the most comprehensive handbook in English; RHYS CARPENTER, *Folktale, Fiction, and Saga in the Homeric Epics* (1946, reissued 1974), a lively comparative account; JOSEPH FONTENROSE, *Python: A Study of Delphic Myth and Its Origins* (1959, reprinted 1980), a massive comparative account with full bibliography; MICHAEL GRANT, *Myths of the Greeks and Romans* (1962, reprinted 1986), discussion of chief myths and their subsequent history; ROBERT GRAVES, *The Greek Myths,* 2 vol. (1955, reissued 2 vol. in 1, 1988), a comprehensive account; JOHN POLLARD, *Helen of Troy* (1965), a popular account of the Trojan saga; PETER WALCOT, *Hesiod and the Near East* (1966), a discussion of Oriental origins of Greek myth; G.S. KIRK, *Myth: Its Meaning and Functions in Ancient and Other Cultures* (1970), a comprehensive critical account; and ANNE G. WARD *et al., The Quest for Theseus* (1970), a full, illustrated account. (A.W.H.A.)

Roman religion: General works include R.M. OGILVIE, *The Romans and Their Gods in the Age of Augustus* (1969), a short account; H.J. ROSE, *Ancient Roman Religion* (1948), a standard work; W. WARDE FOWLER, *The Religious Experience of the Roman People, from the Earliest Times to the Age of Augustus* (1911, reprinted 1971); KURT LATTE, *Römische Religionsgeschichte* (1960, reissued 1976); MARTIN P. NILSSON, *Geschichte der griechischen Religion,* vol. 2, *Die hellenistische und römische Zeit,* 3rd ed. (1974), with a rich bibliography; GEORG WISSOWA, *Religion und Kultus der Römer,* 2nd ed. (1912, reprinted 1971), a basic collection of material; ROBERT E.A. PALMER, *Roman Religion and Roman Empire* (1974); and RAMSAY MACMULLEN, *Paganism in the Roman Empire* (1981). Special periods and subjects are treated in RAYMOND BLOCH, *The Origins of Rome* (1960); MICHAEL GRANT, *Roman Myths* (1971, reissued 1984); H. WAGENVOORT, *Roman Dynamism: Studies in Ancient Roman Thought, Language, and Custom* (1947, reprinted 1976; originally published in Dutch, 1941); MAURO CRISTOFANI (ed.), *Dizionario della civiltà etrusca* (1985); AGNES KIRSOPP MICHELS, *The Calendar of the Roman Republic* (1967, reprinted 1978); W. WARDE FOWLER, *The Roman Festivals of the Period of the Republic: An Introduction to the Study of the Religion of the Romans* (1899, reissued 1969); INEZ SCOTT RYBERG, *Rites of the State Religion in Roman Art* (1955); ALAN WARDMAN, *Religion and Statecraft Among the Romans* (1982); DUNCAN FISHWICK, *The Imperial Cult in the Latin West: Studies in the Ruler Cult of the Western Provinces of the Roman Empire,* vol. 1 in 2 parts (1988); FRANZ CUMONT, *The Oriental Religions in Roman Paganism* (1911, reprinted 1956; originally published in French, 1906); LILY ROSS TAYLOR, *The Divinity of the Roman Emperor* (1931, reprinted 1981); JOHN FERGUSON, *The Religions of the Roman Empire* (1970, reissued 1985); A.D. NOCK, *Conversion: The Old and the New in Religion from Alexander the Great to Augustine of Hippo* (1933, reprinted 1988); MICHAEL GRANT, *The Climax of Rome: The Final Achievements of the Ancient World, A.D. 161–337* (1968); E.R. DODDS, *Pagan and Christian in an Age of Anxiety: Some Aspects of Religious Experience from Marcus Aurelius to Constantine* (1965); ROBERT C. SMITH and JOHN LOUNIBOS, *Pagan and Christian Anxiety: A Response to E.R. Dodds* (1984); and ARNALDO MOMIGLIANO (ed.), *The Conflict Between Paganism and Christianity in the Fourth Century* (1963). See also MICHAEL GRANT and RACHEL KITZINGER (eds.), *Civilisation of the Ancient Mediterranean: Greece and Rome,* 3 vol. (1988), especially the essays in vol. 2. (M.Gr.)

Hellenistic religions: The most useful cultural and political history containing valuable discussions of controversial issues with full bibliography is ROBERT COHEN, *La Grèce et l'hellénisation du monde antique,* new ed. (1948). W.W. TARN, *Hellenistic Civilisation,* 3rd ed. rev. by TARN and G.T. GRIFFITH (1952, reissued 1975); and M. ROSTOVTZEFF, *The Social & Economic History of the Hellenistic World,* 3 vol. (1941, reissued 1986), remain the standard English works. KARL PRÜMM, *Religionsgeschichtliches Handbuch für den Raum der altchristlichen Umwelt: Hellenistisch-römisch Geistesströmungen und Kulte mit Beachtung des Eigenlebens der Provinzen* (1943, reissued 1954), is indispensable for its rich bibliography. The magnificent encyclopaedia now in progress, *Reallexikon für Antike und Christentum* (1950–), will be, when completed, the best single resource for the study of Hellenistic and early Christian religion.

Important general interpretations include PAUL WENDLAND, *Die hellenistisch-römische Kultur in ihren Beziehungen zu Judentum und Christentum,* 4th enlarged ed. (1972); HAROLD R. WILLOUGHBY, *Pagan Regeneration: A Study of Mystery Initiations in the Graeco-Roman World* (1929, reprinted 1974); A.J. FESTUGIÈRE, *L'Idéal religieux des Grècs et l'Évangile* (1932, reissued 1981), and *Personal Religion Among the Greeks* (1954, reprinted 1984); ERWIN R. GOODENOUGH, *Jewish Symbols in the Greco-Roman Period,* 13 vol. (1953–68); SAMUEL K. EDDY, *The King Is Dead: Studies in the Near Eastern Resistance to Hellenism, 334–31 B.C.* (1961); ARNOLD TOYNBEE (ed.), *The Crucible of Christianity: Judaism, Hellenism, and the Historical Background to the Christian Faith* (1969); and LUTHER H. MARTIN, *Hellenistic Religions: An Introduction* (1987). In addition to these works (all of which contain full bibliographies), see the individual volumes in the important series, *Études préliminaires aux religions orientales dans l'Empire romain.* (J.Z.S.)

Human Evolution

Human beings, extant and extinct, comprise the zoological family Hominidae; and the single living human species, *Homo sapiens,* is one of some 200 species of the order Primates, in turn one of 20 orders constituting the vertebrate class Mammalia. Among the past and present diversity of primates, hominids have long been recognized as having the closest resemblances, and hence affinities, to the African great apes (pongids); thus, in 1863 the British biologist T.H. Huxley noted in *The Evidence as to Man's Place in Nature* that "whatever system of organs be studied . . . the structural differences which separate Man from the Gorilla and the Chimpanzee are not so great as those which separate the Gorilla from the lower apes [monkeys]." Various methods for the comparative evaluation of genetic character states have both repeatedly confirmed and measured in some detail the very close proximity of the extant African apes and modern *Homo sapiens.* All such findings are congruent with a common origin of apes and Hominidae, within the African continent, which took place some five to six million years ago.

Three major areas are generally recognized within the subject of human evolution: primatology, which has as its major focus the biological and behavioral aspects of nonhuman primates; human paleontology, which is concerned with the recovery, description, and evaluation of the fossil evidence for hominid evolution; and paleoanthropology, which encompasses interrelated investigations into the biological and behavioral evolution of Hominidae. In addition, five major areas of research can be identified in human evolutionary studies: the origins of Hominidae, adaptation and diversification of the genus *Australopithecus,* the origins of the genus *Homo,* the emergence of *Homo erectus* and subsequent hominid occupation of Eurasia, and the origins and dispersals of premodern and modern *Homo sapiens.*

Investigations of hominid origins are variously concerned with diverse comparative studies of extant higher primates and humans, as well as the search for ancestors in the fossil record. Pongids and hominids show a diversity of contrasting adaptations that evidently reflect their evolutionary divergences and which thus require explanation. Moreover, although markedly different from the Asian pongid (orangutan), the African pongids (gorilla and chimpanzee) differ from one another both structurally and behaviorally. The roots of Hominidae have been traced to at least four million years ago, and possibly to some five million years ago. The rarity and fragmentary condition of the few oldest known specimens, however, do not reveal critical aspects of hominid adaptation, such as modifications in trunk and lower-limb structure. Hence, the details of hominid origins remain unknown and the subject of lively debate and substantial speculation. The ancestral stock of extant African apes and of hominids also remains unknown, in large part a reflection of the paucity of fossil-bearing localities in the five- to 10-million-year time span. In the absence of a fossil record, structural and other adaptations have been projected back as an ancestral condition from living descendant species; but this is a very risky procedure that dismisses morphological transformation and adaptation and assumes stasis without complementary confirmation.

The oldest definitely known hominids are attributed to the extinct genus *Australopithecus.* The genus speciated substantially, producing distinct and, in some cases, possibly convergent lineages. At least four species (*afarensis, africanus, robustus, boisei*) are commonly accepted, and two more (*aethiopicus, crassidens*) are recognized by some workers on morphological grounds. All species of the genus originated in the Pliocene epoch (5.3 to 1.6 million years ago), and the genus apparently became extinct in the Early Pleistocene (about 1.6 million to 900,000 years ago); its distribution is unknown outside the African continent. The oldest and most primitive species is *A. afarensis,* and most workers believe it to be ancestral to succedent species. Although they exhibited some fundamental hominid adaptations (bipedalism, reduction of anterior dentition, exploitation of nonforested habitats), most or all australopithecine species remained primitive in terms of growth and maturation, brain size and proportions, dietary adjustments, and complexity of cultural behaviour. *Homo* coexisted with the later, so-called "robust," australopithecines—*robustus* (and possibly *crassidens*) in southern Africa and *boisei* in East Africa—although the adaptations enabling such coexistence are scarcely understood, and it is generally thought that an australopithecine species was ancestral to *Homo.*

The recognition and suitable definition of the genus *Homo* and its initial representatives has been a persistently troublesome problem. There have been no formal diagnoses, and the few characterizations offered suffer from both lack of definitive character states and inclusiveness. The problem has been exacerbated as the hominid fossil record has expanded, particularly in respect to specimens dated to the end of the Pliocene epoch that lack distinctively *Australopithecus*-like characteristics. The first such specimens, found in the early 1960s in the Olduvai Gorge of Tanzania, were designated *Homo habilis.* Further remains of both comparable and greater age were subsequently recovered from northern Kenya and southwestern Ethiopia. Although a single initial *Homo* species (*H. habilis*) was originally proposed, this perspective has been criticized by some workers as simplistic because of the substantial variability of the fossil finds; accordingly, it is entirely possible that two contemporaneous and even sympatric species may have existed in the Late Pliocene. Coincident with the appearance and subsequent presence of such hominid(s) are various traces of associated culturally patterned behaviours. These include evidence of natural but transported and accumulated stone, flaked-stone artifacts, and occasional associated mammal (and other) skeletal parts, all of which indicate the exploitation and utilization of animal resources; the repeated utilization and occupation of particular locales; and the expanded employment of natural resources in conjunction with technological capabilities and requirements. Such biological and behavioral adaptations are believed by many workers to reflect major transformations and reorganizations in hominid phylogeny, perhaps consequent upon the initial appearance of genus *Homo.*

The fossil record in sub-Saharan Africa affords evidence of the appearance of another, more derived (*i.e.,* more evolved) species of *Homo—Homo erectus*—at the beginning of the Pleistocene epoch. At several localities in East and southern Africa, the species occurs sympatrically with the "robust" australopithecines. Less ancient occurrences are also known from northwestern Africa. The initial occupation of Eurasia by hominids appears to postdate such an antiquity, and it is generally inferred that the first Eurasian hominids were dispersals from an African source, perhaps between 1.5 and one million years ago. *Homo erectus* was first and, for a long time, best known from fossil finds in Southeast and East Asia. The fossil occurrences there range in age from approximately 1.6 million to 250,000 years. Although initial hominid occupation in Europe was probably at least as early, no human skeletal remains are known from the most ancient times, and those that have been found—dated to between 500,000 and 300,000 years ago—do not represent *H. erectus* but rather a form of *H. sapiens* that has been labeled "archaic." The initial and subsequent penetration of hominids from the lower to middle latitudes occurred as the amplitude and intensity

of glacial–interglacial climatic cycles was increasing. The extent to which there were attendant, and perhaps correlative, changes in human biology and in behaviour have remained the subjects of substantial research and much controversy.

Traditionally, the tendency among students of hominid evolution was to attribute premodern human fossil finds to one or more extinct species, and sometimes even distinct genera, but as efforts increasingly have been directed at seeking congruence between developments in evolutionary biology and in the state of the hominid fossil record, substantial revisions in the classification of Hominidae have emerged. A variety of premodern human finds of both late Middle Pleistocene and early Late Pleistocene ages came to be subsumed within the species *H. sapiens* and were only further distinguished below the species level. This varied and increasingly large sample of post-*erectus* hominids came to be regarded as archaic *H. sapiens,* as distinguished from anatomically modern humans. This is largely a consequence of the mosaic of morphological features represented in the substantial variations among such materials, which exhibit primitive versus advanced, or derived, features.

If early African *H. erectus* constituted the source of subsequent hominids, then it would appear that evolution proceeded quite differently in major geographic areas. For example, *H. erectus* is characteristic of the Asian Middle Pleistocene, where it is also long persistent and distinguished by its own group of singular, derived features. In Africa derivatives of early *H. erectus* are known, as are some transitional examples linking *H. erectus* to archaic *H. sapiens.* Skeletal parts of the earliest hominid occupants of western Eurasia are not known. The first examples from this area occur well into the Middle Pleistocene, and there is a range of variation in the specimens from strongly and partially *erectus*-like to incipiently Neanderthal-like, passing ultimately into the well-known and widely distributed Neanderthal peoples. The Neanderthals were for many years treated as a distinct species (*Homo neanderthalensis*), but they were subsequently subsumed as an archaic subspecies of *H. sapiens.* An increasingly substantial body of evidence has been accumulated, however, which suggests that a return to the older position is probably warranted. In western Europe, at least, there is increasing evidence of the contemporaneity of the last Neanderthal peoples with those early modern populations that have come to be known as Cro-Magnons. For these reasons, and in order to recognize and express differing degrees of derivation, further taxonomic evaluation and distinction of these archaic *H. sapiens* specimens is required, which will doubtless include the recognition of additional subspecies of both the *H. erectus* and *H. neanderthalensis* groups in western Eurasia and also corresponding taxonomic reassessment of various African Middle Pleistocene samples. In East Asia the existence and morphology of archaic *H. sapiens* has been well established, but the extent to which this form was a contemporaneous or a succedent replacement for the late *H. erectus* populations has not been firmly resolved.

The roots of anatomically modern humans have long been a puzzle to students of human evolution and hence the source of much speculation and debate. Nonetheless, several developments have caused renewed interest in the problem from different perspectives. First, there has been the recognition of substantially greater relative and absolute ages for the archaeological industries of the African Middle Stone Age (about 200,000 to 40,000 years ago) and, correspondingly, some associated modernlike human skeletal parts. A series of such sub-Saharan occurrences has been identified within earlier and later segments of this time span. Second, there has been the recognition and broad acceptance of early modern human (often called "Cro-Magnoid") populations in western Asia that were distinct from, and considerably older than (*i.e.,* 90,000–100,000 years ago), the known Neanderthals from that area and also much older than the European Cro-Magnon peoples who were widespread in Europe by some 30,000–35,000 years ago. Third, there has been the increasing availability of comparative genetic data on degrees of affinities of modern human populations. The biochemical systems of Asian and European populations appear to be more similar to each other than those of either group are to African populations; thus, Asians and Europeans may have shared a common ancestry some 40,000 years ago and a common ancestry with African populations almost three times as long ago. Moreover, investigations of human mitochondrial DNA reveal two facts: that the variation among modern human populations is small compared, for example, with that between apes and monkeys, which points to the recency of human origin; and that there is a distinction between African and other human mitochondrial DNA types, suggesting the substantial antiquity of the African peoples and the relative recency of other human populations. (F.C.H.)

The major focus of this article is on the physical evidence—both fossil and lithic—for human evolution and on the interpretation of this evidence. For a detailed discussion of evolutionary theory, see EVOLUTION, THE THEORY OF; for a detailed discussion of the evolution of human behaviour and culture, see PREHISTORIC PEOPLES AND CULTURES. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 411 and 412, and the *Index.*

The article is divided into the following sections:

GENERAL BACKGROUND TO HUMAN EVOLUTION

Evolutionary process in man

Evolution may be defined as change in the genetic composition of a population through time (see also GENETICS AND HEREDITY, THE PRINCIPLES OF). It may be thought of as progressing on two levels: the production of variation and the selection of those gene combinations most fit in a particular environment. It would be wrong to consider that evolution is "caused" by any one factor. Indeed, the error of many earlier theories of evolution was neither intrinsic nor factual but lay in the consideration of evolutionary change solely in the light of a single process. While the modern theory of evolution relies heavily on the Darwinian concept of natural selection, many other factors unknown in Darwin's day have enlarged and enhanced the knowledge of evolutionary process. The modern synthesis of evolutionary theory is a composite approach that views evolutionary process in the light of different scientific disciplines such as paleontology, biology, biochemistry, ecology, and many others, all of which have made significant contributions. The synthesis of fact and theory from these various fields has been the major contribution of evolutionary biology to 20th-century science.

GENERAL CONSIDERATIONS

Darwinian evolution

With the publication of *On the Origin of Species by Means of Natural Selection,* Charles Darwin in 1859 showed conclusively that species evolved and were not immutable over time. This revolutionary idea permitted an explanation of the fossil record that did not need to invoke the biblical story of the Flood or the view that all extinct animals and plants had perished as a result of this one global catastrophe. It became possible to compare modern and fossil animals and to construct lineages through time that documented the changes that had occurred, and the distribution of fossil forms began to take on new significance. It became apparent that assemblages of fossils betray climatic preferences at any given time and climatic change through time. Another 12 years elapsed before Darwin applied his theory of evolution—and its mechanism, natural selection acting upon a pool of normal biological variation—to the case of man; the delay more likely was because of lack of fossil evidence than lack of courage.

The so-called Darwinian tautology, "The survival of the fittest is the survival of those best fitted to survive," gives an insight into the adaptations of living organisms that lead to an increase in their chances of survival and of leaving more offspring than their rivals. The closer the adaptation to the environment, the greater the chances of survival. This pathway leads to specialization: fish need water in which to swim, birds need wings with which to fly, koalas need eucalyptus leaves to eat—nothing else will do.

This approach to survival has its advantages but also its drawbacks. Should the environment change suddenly, those who have gambled on specialization may lose, while those who have retained a generalized form and remained adaptable can adjust to the new situation and survive. On the whole, the order Primates, which contains humans and their ancestors, has retained this approach, an evolutionary flexibility that has enabled primates to respond to change when it has arisen.

Dating fossils. The period of time involved in human evolution is at least three to four million years and probably much longer. During this time the world has undergone a number of climatic changes, including glaciations and warm periods, which can be detected by studying the geologic record. The variety of dating methods that are available to paleontologists makes it possible to place fossils in time and succession. The fossils can also be assembled into faunal groups that provide ecological information; from such information, coupled with the analysis of skulls, bones, and teeth, a picture begins to emerge of the creatures in their world or, as evolving lineages are sampled, in successive worlds. No such opportunity was granted to Darwin.

Two general categories of dating methods exist: relative and absolute dating.

Relative dating. Before the age of a fossil can be determined, it is first necessary to establish that it is contemporary with the deposit in which it was found and not a more recent, intrusive burial in an older deposit. Often, contemporaneity can be determined by the fossil's association with other fossils or rock strata of known age from the same or other sites. This will establish what is called a fossil site's biostratigraphy. Chemical tests for fluorine and nitrogen in fossilized bone should show equivalent amounts throughout an assemblage (if it is contemporaneous), and the newer technique of X-ray microanalysis may establish an elemental "fingerprint" that is typical of a given location or stratum. An absolute method employing the decay of uranium also is useful for establishing contemporaneity.

Absolute dating. Once contemporaneity has been determined, the age of a fossil in years can be approached using several methods that collectively are called absolute, or chronological, dating. These methods rely on the fact that a number of radioactive isotopes (such as uranium) are known to decay into daughter products at known constant rates. Measuring the amount of uranium that has decayed into thorium, for example, can be used to determine the age of some rocks if uranium was laid down with the original deposit.

Carbon-14 dating

Probably the most popular and best-known radiometric method is called carbon-14 (radiocarbon) dating. Carbon-14, a naturally occurring isotope of carbon-12, is absorbed by living organisms at a known rate. Absorption ceases when an organism dies, and its age can be determined by measuring the ratio of carbon-14 to carbon-12; the method is effective up to about 60,000 years ago and perhaps up to 100,000 years ago using a modern particle accelerator. Another highly effective method is called potassium–argon dating. Radioactive potassium decays into argon gas, which becomes trapped in volcanic rocks as they cool. The amount of argon relative to radioactive potassium can be measured to yield reasonably accurate dates on volcanic-rock samples that are more than 250,000 years old.

Other dating techniques include the fission-track method, in which the tracks made by uranium atoms undergoing spontaneous fission can be measured; this method is useful in bridging the gap between the carbon-14 and potassium–argon methods. Changes in the Earth's magnetic polarity that are preserved in rock strata also provide an accurate dating method. An even more recent dating method employs the principle of electron paramagnetic (or spin) resonance, a form of microwave absorption spectroscopy in which the magnetic moment induced by the self-rotation of negatively charged electrons can be detected.

Mosaic evolution. Evolution has been defined as change in the modal human phenotype, a genetic idea that suggests a shift in the makeup of the average man or woman through time and through successive generations. It is a process that is continuous but at varying rates in response to environmental pressure and natural selection. The idea that not only do differing populations evolve at differential rates but also that individual parts of the body may evolve differentially is called mosaic evolution; recognizing and applying this phenomenon has greatly expanded the understanding of human evolution. The makeup of individuals is determined by their genetic endowment, so that the very wellspring of evolution will be found in that variability that defines all humans as individuals.

GENETICS AND MAN

The ultimate source of all new genetic variation is change in the genetic material itself. Genetic information is passed from parent to offspring through the complicated protein molecule known as deoxyribonucleic acid (DNA). The DNA is carried on long strands called chromosomes. Each species has a characteristic number of these strands—in humans the number is 46. These 46 chromosomes com-

prise 23 pairs. The DNA molecules in similar positions along paired strands will code for the same feature, and, in the chemical relationships between these positions (or genes), one gene of each pair will generally be dominant, and the other will be recessive. It is usual for the activity of a gene on one strand to be masked or inhibited by its counterpart on the other strand. An important exception to this is in the determination of coloration patterns, in which the genes may appear to blend, producing colour patterns different from those of the parents.

Types of mutation

Mutation. Alteration of the genetic material is called mutation and can be of two types: point mutations, in which the molecular composition of a discrete location on the chromosome strand can be changed by a chemical process or the physical interaction of a particle (perhaps resulting from radioactivity) with the DNA molecules that make up the chromosome at that point, and mechanical mutations, in which the gross structure of the chromosome is altered, often in the process of cell reproduction. Point mutation is well documented in human populations; for example, normal hemoglobin is formed of a long chain of amino acids; if the genetic code for one of these amino acids, glutamic acid, is changed, the amino acid is replaced by another amino acid, valine, and the abnormal hemoglobin responsible for sickle-cell anemia results. This disease is of considerable importance in some human populations and will be discussed in more detail below. Mechanical mutations do not involve chemical changes in the DNA but are physical alterations involving the body of the chromosome itself; these abnormalities usually occur during cell reproduction. One such process involves the exchange (called crossing-over) of material between two adjacent paired chromosomes. Other types of chromosomal mutations can be due to the inversion or translocation of chromosomal material, its loss or deletion, or the nondisjunction of chromosome pairs.

While the ultimate source of all intrinsic genetic change must occur through mutation, the vast majority of these alterations result in the death of the individual or so reduce his fitness that he suffers "genetic death" (*i.e.,* fails to reproduce) even though he may live to great age. The greatest source of variability is simply a change in the frequency of existing genes under various environmental and, in some cases, social stimuli. The genes themselves contain a vast reservoir of variability, only part of which can be realized in any single population. Here an important distinction must be made between the genotype—the chemical composition of the genes—and the phenotype—the external appearance—of the individual. Similar genotypes can in certain situations give quite different phenotypes, presumably because only part of the genes' activity occurs at any one time. This phenotypic flexibility may be of considerable importance in the origin of certain racial variations in living groups, but its long-term significance in evolutionary process is not yet understood.

Most variability arises, then, as existing genetic material is recombined or reshuffled into new patterns via sexual reproduction. Alterations in gene frequencies will occur mainly under pressure of selection. Besides the reservoir or variability that any sexually reproducing local population contains, another important source of genetic potential can be derived from interbreeding with nonlocal groups. This is sometimes called gene flow, and, although there may be a temporary disruption of the genetic equilibrium, once a balance is again obtained the hybrid population may be more fit than the parent generation. Heterosis, or hybrid vigour (the phenomenon whereby an organism that is the result of crossbreeding between parents of different genetic populations shows increased fitness and general vitality), is well documented for some human groups, most notably the people of Pitcairn Island, who are ultimately derived from crosses between Tahitian women and the British sailors from the HMS *Bounty,* groups whose ancestors were genetically separated by distance for thousands of years.

Hybrid vigour

Observations on the occurrence of genetically derived traits in fossil hominids can be used to suggest earlier patterns of gene flow. Such observations have been made with regard to *Homo erectus,* and the suggestion has been made that genetic exchange was occurring between widely spaced members of this group during the Middle Pleistocene (about 900,000 to 130,000 years ago). It has been pointed out by a number of workers that the approximately contemporaneous mandibles (lower jawbones) found at the Ternifine site in Algeria and the mandibles found near Peking show extreme similarities; the great similarities between the Peking femurs (thighbones) and that of Olduvai Hominid 28 found in the Olduvai Gorge of northern Tanzania have also been noted, as have the similarities between the pelvis of OH 28, the hominid pelvis designated KNM-ER 3228 from the Koobi Fora site in northern Kenya, and the Arago XLIV hominid pelvis from Tautavel in southern France. A reasonable explanation of this similarity is that migratory hunting patterns had brought many groups of *H. erectus* into contact and that exogamous (marrying outside the tribal group) breeding patterns had resulted in the widespread occurrence of certain traits. These similarities are very likely too great and consistent to have resulted from separate evolution along parallel lines in isolation; and, indeed, the degree of similarity seen in the available material makes it extremely unlikely that long-term isolation was a factor in human evolution after the early Middle Pleistocene.

Natural selection. Genetic variation derives from several sources and provides only the raw materials of evolution; random variation can only have a disrupting effect on genetic equilibrium unless deleterious combinations of genes are eliminated and advantageous ones are preserved. The process by which this occurs is called natural selection. Although Darwin is credited with the first full statement of the theory of evolution through natural selection, the origins of the concept are deeply rooted in European thought and may be traced to the early 17th century. Indeed, the practices of artificial selection of domesticated plants and animals, which so influenced Darwin, can be traced at least as far as the Romans and perhaps even to Neolithic (New Stone Age) times. The theory of natural selection can be stated as follows: All living things vary and reproduce themselves many times, yet the number of a given group tends to remain constant. Therefore there is a competition for survival, and only those most adapted to external conditions survive. In essence, the idea of natural selection is statistical; those members of a population who are most evolutionarily fit are those who will leave the greatest number of offspring. These will not be the only members of the population to leave offspring, but it is probable, in a statistical sense, that they will leave more living descendants in the long run than the less well adapted members of the population. Fitness therefore refers to a population's ability to cope successfully with a particular environment at a particular time; it is tied to time and place in an absolute way. The factors that help determine the course and direction of evolution are many: predation and disease, migration and conflict, behaviour and temperament, competition for breeding space and mates, and competition for living space and food. Purely physical factors in the environment are no less important: stability or instability of the climate; solar radiation; natural disaster; pollution of the soil, water, and air—all will have their effect and take their toll on living groups. Evolution is not, therefore, something that has occurred and been completed. Any population is constantly evolving, assimilating changes and variations in its gene pool in response to stimuli from a large number of sources—some that are recognized and some that are not.

Evolutionary fitness

ADAPTATION AND GENETIC CHANGE

There are a number of ways in which a population can adapt to the changing world in which it lives; under the heading of natural selection, it is possible to define several mechanisms that help maintain a viable relationship between that population and the environment of which it is a part. In a seeming paradox, natural selection can work to maintain the status quo in a changing environment while at the same time drawing on new or existing material in the gene pool to meet the demands of the environment in a new way. Natural selection is then both conservative and dynamic, with both mechanisms work-

ing to achieve the same end result, that of adaptation to the environment. Conservative, or stabilizing, natural selection basically works to eliminate detrimental genetic effects or genes harmful in a particular situation; it tends to reduce variation in the gene pool. It would appear that this aspect of natural selection has become less effective in man with the development of modern medical research and practice. Chronic and often lethal diseases caused by genetic conditions, such as diabetes mellitus, hemophilia, and phenylketonuria, are certainly more common now than several decades ago because of modern medical care, since persons with these diseases live and produce children with the disease, although they would have died or failed to reproduce under less advanced medical techniques. These individuals, however, are not less fit in a Darwinian sense than the "normal" members of the population as long as their environment includes this medical care; it is only in the absence of this care that they are less fit.

Homeostatic change. A nongenetic correlate of conservative natural selection is homeostasis. This term refers to adaptive physiological cultural flexibility that is genetically based, and it means the retention and preservation in a population of its internal equilibrium in the face of disruptive external environmental conditions. By drawing on this source of variation, a population may adapt its physiology or behaviour in response to new environmental demands without actual changes in the gene pool; because it is an immediate response, it may enhance a group's survival potential. One type of homeostatic response has been called acclimatization; the efficiency and scope of this process are genetically based, yet the full range of responses is seldom, if ever, activated. Because man occupies a wider range of habitats and is therefore exposed to more extreme conditions than any other species, the necessity for broad responsive ability to changes in environmental conditions such as climate is obvious. Homeostatic responses merge undetectably with both dynamic and conservative natural selection and cannot, indeed perhaps need not, be separated from them for most purposes. The extent of homeostatic response is unique in man, not only in the diversity of the physiological responses that deal with the extremes of the environment but also because of the added dimension of culture as an adaptive mechanism. It is this aspect, culture, that allows humans not only to fully inhabit and utilize a wide range of environments but also to alter these environments to their own ends. The diversity of the ways in which humans deal with environmental extremes can be seen in the various mechanisms (skeletal, physiological, and cultural) used by people in cold climates.

Acclimatization

Blood types, abnormal hemoglobin, and disease. Diversifying, or dynamic, natural selection is one of the important and basic processes by which evolutionary changes occur. Under conditions of changing environmental pressure, advantageous genotypes will be assimilated into the gene pool, and those individuals within the population who have superior fitness will leave more numerous offspring than those without. One such adaptation is the maintenance of a potentially lethal gene for an abnormal hemoglobin (hemoglobin S) in populations that are exposed to malaria. Sickle-cell anemia and several related anemias are genetically based blood disorders that can kill an individual who inherits a gene for the condition from both parents (homozygous condition); those with only a single abnormal gene (heterozygous condition) will demonstrate the disease under certain conditions but are less affected and usually do not die of it. In certain areas of tropical Africa, however, as well as around the Mediterranean and in regions of East Asia, individuals without one gene for the abnormal hemoglobin can die or become seriously ill from malaria before they have the chance to reproduce. The presence of the abnormal hemoglobin therefore confers protection in certain environments, while outside these areas the abnormal gene has no adaptive significance and will be selected against. The abnormal hemoglobin is virtually absent in nonmalarial areas except in the descendants of people from such areas.

Lethal genetic disorders

Other examples of diversifying selection are to be found in the distribution patterns of the ABO blood group system (see BLOOD: *Blood groups*). Although the data is incomplete, it would appear that individuals with certain blood types are more susceptible to certain diseases and that, therefore, in areas where these diseases are common, these blood groups will be selected against. Suggestions have been made that blood group O individuals are less susceptible to syphilis, group A to plague, and group B to streptococcal infections; fieldwork seems to support statistical correlations between certain types of disorders of the gastrointestinal tract and some ABO blood groups. For example, a higher frequency of duodenal and gastric ulcers in group O and a higher incidence of cancer of the stomach of group A have been demonstrated.

Random genetic drift. Genetic drift is another mechanism by which evolutionary changes may occur. Drift can be defined as the apparently random variation of certain gene frequencies under special conditions of small population size or of isolation or both. Also called the Sewall Wright effect and non-Darwinian evolution, it has been the subject of considerable controversy. From these discussions it emerges that many of the features previously thought caused by drift are now known to have been the result of previously unrecognized natural selection. Central to the definition of genetic drift is the assumption that genes can be neutral and without effect in terms of evolutionary adaptation, and it is this assumption that has led to the greatest controversy. Some workers have argued that 5 to 10 percent of all mutations are selectively neutral and may be maintained in the gene pool without effect. Others have asserted that, because genetic systems have complex chemical interrelationships, with each gene contributing something to the finished product, it is unlikely that a single gene or group of genes could be totally without effect. This latter position recognizes that, while certain observable features may have little apparent effect on fitness, the genes responsible may be inextricably associated with other genetic features that do have important fitness correlations.

The near absence of blood group B in the American Indian has been attributed to genetic drift. However, B-type blood reaches its highest frequency in some Mongoloid groups thought to be ancestral to the early inhabitants of the New World. This and the probability that many incursions from Asia must have occurred in the Late Pleistocene epoch (35,000–10,000 years ago) make drift a somewhat untenable explanation of the lack. A more likely explanation is that holders of B blood were faced with some as yet undefined environmental hazard in the New World that made blood type B a genetic liability and it was subsequently lost through selection.

CULTURALLY PATTERNED BEHAVIOUR AND EVOLUTION

Biological or organic evolution takes place as a result of the process of natural selection affecting a pool of normal variability within a living population. This genetically based variability principally involves the physical characteristics of the individuals who make up the population, such as size, shape, coloration, or susceptibility to disease. For many years these alone were thought to be the only mechanisms of evolutionary change. Animal behaviour was studied as a discipline on its own, and human behaviour and human cultures were studied similarly in relative isolation. It has come to be realized, however, that behavioral traits in animals and in humans may under certain circumstances have profound effects on breeding patterns, mate selection, and the survival of offspring. If this is the case, then these behavioral and, in man, cultural attributes will affect the course of organic evolution through their modification of the selective processes operating on the gene pool. The study of these behavioural and cultural effects in relation to evolution has been termed sociobiology.

Natural selection can be shown to operate at a number of levels within populations, but the ultimate level of selection may be at the molecular level—that of the genes themselves. Perfect replication occurs over millions of years so that while individuals and groups of individuals may come and go, the genes march on forever as seemingly immortal. They have been termed "selfish genes" or "survival machines" that are concerned only with their

own survival and are spread, as it were, by driving the bodies that carry them to increase and multiply. Given a stable environment and gene pool, this strategy is said to be evolutionarily stable; a new gene entering the pool that could influence the behavioral strategy would be selected against.

Most animal behaviour is directed toward the reproductive advantage of the individual. Occasionally among social insects such as ants or bees, however, workers are seen not to take part in reproduction, a behaviour called altruistic with respect to the colony and its future. In other species, different forms of altruistic behaviour can be observed, such as "aunts" sharing the rearing of young; in all cases, such behaviour is paradoxical from the viewpoint of individual selection. It may be, however, that the donor and the recipient of this type of behaviour share a larger number of genes or gene types than previously had been believed, so that the "sacrifice" may help the genes held in common with the recipient to survive and multiply. If this is true, then the altruism of close relatives is given purpose and kin selection is an effective strategy.

Altruistic behaviour

There are those who believe that these ideas can be transferred from animal behavioral studies to the more complex realm of human behaviour and culture, but the study of sociobiology is not without controversy. For example, the relationships between mothers and infants are strong in all mammals; in some mammals and birds pair-bonding is lifelong, but in others, including primates, are found males with harem groups and altruistic behaviour by "aunts" and "uncles" toward "nephews" and "nieces." If comparisons are made with human kinship systems—the stuff of social anthropology—the typical pattern is that of monogamy (single pair-bond), with less frequency of polygyny (multiple-female, single-male groups) and, occasionally, even polyandry (multiple-male, single-female groups). Unlike the animal system, however, the basis of the human system may be cultural or even economic, and there appears to be no correlation between cultural kinship (*i.e.,* shared behaviour) and genetic kinship (shared genes).

The relationship between organic evolution and cultural evolution is complicated, since each form appears to have differing mechanisms. Culture as a uniquely human attribute seems bound to affect the totality of humanity and through this totality the patterns and results of human reproductive behaviour; this, in turn, must affect the human gene pool and the selective processes that act upon it, which is the essence of human evolution.

MAN'S CONTINUING EVOLUTION

The question of man's continuing evolution may be posed on two levels, because it is possible to define two distinct yet interrelated levels of evolutionary change: the first can be called phyletic evolution and the second, phenetic evolution.

Phyletic evolution. This term refers to cumulative and important changes in the population gene pool that lead eventually to speciation (separation into new species) and the higher levels of taxonomic differentiation. It is phyletic evolution that is usually identified in the fossil record. Phenetic evolution is more subtle, and, while it is identifiable in living populations, it is less easy to see in fossil groups. It involves changes of a lower magnitude than phyletic evolution and may be called ecotypic evolution. As with all evolutionary changes, phenetic changes occur in response to environmental stimuli, but it is the consistency of these stimuli through time that determines whether the changes will become permanently impressed on a group's phyletic record or if they will be lost within a relatively short period due to changing pressures on the gene pool. Phenetic evolution is the mechanism underlying the formation of subspecies, races, or varieties.

Levels of evolutionary change

It is difficult to identify any unequivocal evidence of phyletic evolution in man for perhaps the last 250,000 years, and the reasons for this difficulty are not hard to find. Human adaptations to the environment have been as broad and generalized as the ecological niche that humans occupy. As a species man inhabits and uses possibly more of the Earth's surface than does any other species, and the breadth of this niche demands considerable flexibility in the human gene pool. An additional factor in man's at least temporarily arrested phyletic evolution is the intervention of culture between man and his environment. With culture, first in the form of crude tools and perhaps skin clothing and now with a variety of sophisticated technologies interposed between humans and the natural environment, the environment cannot exert pressures on the human species in the same way that it has in the past. A third factor must certainly be the size of the gene pool itself and the relatively extended length of a human generation. Evolutionary changes can occur with considerable rapidity in small populations and in populations that mature and reproduce quickly; but the flow of favourable genotypes, however advantageous, must be extremely slow in man's particular circumstances. In terms of man's total morphological pattern and its component parts, then, no important changes are preserved in the fossil record since late Middle Pleistocene times. The last known evidence of human phyletic evolution concerns the *H. erectus–H. sapiens* transition; this period of transition occupied a considerable period of time, and, because of the incompleteness of the fossil record, scholars can only guess at many of its details. Certainly, *H. sapiens* did not spring fully formed from their *H. erectus* ancestors but, through the process of mosaic evolution, crossed the sapient threshold at varying times in the development of different functional complexes. The resultant combination of *H. erectus* and *H. sapiens* features is well demonstrated in a number of specimens, notably Omo II, Steinheim, and Vértesszőllős (see below *Homo sapiens*).

Phenetic evolution. The apparent arrest of man's phyletic evolution should not suggest that the environment will never affect human development again; the observed fact of continuing human phenetic evolution clearly demonstrates the possibility of further phyletic evolution. Some of the clearest evidence that man is still responding to environmental stimuli comes from studies on the distributions of ABO blood groups; this and the importance of sickle-cell anemia in this context have already been described. Further evidence for phenetic evolution is to be found in the analysis of certain general morphological patterns, such as Bergmann's rule, which states that within a polytypic warm-blooded species the body size of a subspecies usually increases with decreasing temperature of its habitat, and Allen's rule, which states that in warm-blooded species there tends to be an increase in the relative size of protruding organs such as the ears and tail with increasing temperature of the habitat.

Disease is an especially acute natural selective agent, and it is well recognized that many diseases recur in cyclic patterns when their virulence is increased. This may be associated with conditions in the environment favourable to the disease-producing organism (pathogen) or by changes in the organism that suffers the disease (host); plague, tuberculosis, and scarlet fever may demonstrate this sort of pattern. In addition, populations that do not have natural immunity may be decimated by a disease that is relatively mild in nonsusceptible groups. Many Polynesian peoples, for example, have been virtually annihilated by measles. It is possible, therefore, that highly lethal worldwide epidemics could act as potent selective forces; such selection could be caused by some new or previously unrecognized disease or by a new episode of high lethality in an existing one.

Disease as a selection factor

The effects of disease are one example of how further evolution in man may occur. Although clear conclusions with regard to human populations are not yet available, experiments with laboratory animals indicate further possibilities. A number of pathologies, both social and physical, have been induced in laboratory animals under conditions of crowding: infertility, cannibalism, mental aberration, and early death have all been observed. Environmental pollution is also known to seriously damage laboratory animals, but again the application of these results to human populations is not yet defined. Clearly, in order to obtain an understanding of the course of human evolution in a modern context, it is necessary for information to be combined from a wider range of disciplines than was formerly the case. The facts of organic evolution as the overall

mechanism by which man has evolved are no longer in serious dispute; undisputed too is his continuing need for reactive and progressive changes, whether evolutionary or cultural, in response to a changing world.

Man's evolutionary relationships

THE LIVING PRIMATES

Taxonomic classification includes man as a member of the order Primates, which is a part of the class Mammalia. Within the Primates are included such divergent creatures as the Southeast Asian tarsiers, the Madagascan lemurs, the South American monkeys, the African monkeys, the great apes (gibbons, orangutans, chimpanzees, gorillas), and finally, humans themselves (see Table 1). The primates that exist today comprise a remarkable gradational series that links *Homo* anatomically with small mammals of very primitive types. The most lowly representatives of the living primates are the tree shrews, small squirrellike creatures that have a wide distribution in Southeast Asia. So primitive are the tree shrews that many authorities refuse to include them among the primates. But even these authorities would agree that tree shrews must at least be very closely related to the ancestral stock from which the primates in general have derived. In a number of their anatomic characters, the tree shrews show such a close resemblance to undoubted primates (*e.g.*, some of the lemurs) as to amount in certain details to an identity of structure. The lemurs in their more advanced anatomic structure show a mixture of characters that indicate an intermediate position between tree shrews and monkeys. The curious little tarsier (which inhabits Borneo, Sumatra, the Celebes, and the Philippines) is in some respects even more monkeylike. The various types of tailed monkeys represent a still higher grade of organization and through the gibbon, smallest of the apes, are linked with the larger tailless anthropoid apes: the chimpanzee, orangutan, and gorilla.

Table 1: Classification of Man Within the Order Primates

Order	Primates	contained forms: prosimians, monkeys, apes, ape-men, and humans
Suborder	Anthropoidea	monkeys, apes, ape-men, and humans
Superfamily	Hominoidea	apes, ape-men, and humans
Family	Hominidae	ape-men and humans
Subfamily	Homininae	humans
Genus	*Homo*	archaic and modern humans
Species	*erectus* *sapiens*	

The manlike apes

As their name implies, the anthropoid apes are manlike in their anatomic structure. Their brains, although much smaller than a modern human brain, are relatively well developed as compared with lower primates and have the same patterns of convolutions as the human brain has (though in a simplified form); the similarities in many details of the intrinsic structure of the brain of the anthropoid ape are astonishingly precise. These anatomic resemblances in the brain have been found to be correlated with physiological similarities. Thus, the sensory and motor mechanisms fulfill functions so closely reproducing those of the human brain that for experimental studies anthropoid apes have been found to be far more reliable than any other nonhuman mammal in their application to problems of cerebral function in man. Many features of the skull and skeleton of the large apes approximate very closely those of the Hominidae (the taxonomic family that includes living and fossil humans), particularly if account is taken of certain extinct primitive hominids. Some of the structural similarities in the skeleton of the trunk and limbs are in part related to posture, for the chimpanzee and gorilla are capable, at times, of balancing themselves on their hind limbs in a manner that suggests, albeit distantly, the erect posture characteristic of the Hominidae. In their dentition, particularly in the molar teeth, anthropoid apes also show a close resemblance to the Hominidae. Indeed, it is sometimes difficult to determine whether isolated fossil molar teeth belong to apes or hominids, for the distinctions that exist between the teeth of apes and those of humans are in general far less obtrusive when fossil specimens are considered. Many of the muscles of the human body have the same disposition and attachments as those of the anthropoid apes. In the sole of the human foot, for example, are found the same muscles that are used for the mobile functions of the ape foot, even though in man the mobility is not present. The disposition of the abdominal organs in apes corresponds quite closely with that of man, and even in their microscopic details some of the organs of the body show a remarkable resemblance. These examples of anatomic and physiological similarities between the large anthropoid apes and the Hominidae could well be multiplied. Their implications for a real phylogenetic relationship are further supported by reference to similarities in serum protein patterns, immunologic responses, some of the blood groups, parasitic infestation, and susceptibility to certain diseases.

Trends of primate evolutionary development. As diversified as the primates are in terms of appearance, ecological adaptations, and behaviour, through their common ancestry they share a number of distinctive characteristics. Fundamentally they have avoided extreme anatomic specializations and have, for the most part, preserved a generalized morphology, which confers on them a considerable degree of functional plasticity. The order can be defined on the basis of the prevailing trends that have dominated its evolutionary development: a generalized structure of the limbs with the replacement of sharp claws by flattened nails resulting in a grasping hand, the elaboration of the visual sense and a corresponding reduction in the sense of smell, and the progressive development of a large and complicated brain. In the Hominidae these same tendencies have manifested themselves and have progressed further than among other groups of primates. The Hominidae show also a unique specialization of the lower limb; the pelvis and leg have become adapted for support and bipedal propulsion of the body in the erect position, while the foot and toes have lost the prehensility characteristic of the primates in general.

Primate generalized morphology

In examining the development of the primates over a period of perhaps 70 million years, the evolutionary trends that have emerged form a pattern of progressive adaptational changes largely related to arboreal life. Although a number of separate trends can be traced individually, they are largely grouped around several functional and behavioral centres. The very existence of evolutionary trends has

been questioned, and some workers have suggested that they existed only in the mind of the paleontologist. But, if it is true that directed, straight-line, orthogenetic trends do not exist in nature, it is also true that within a single evolutionary lineage only certain types of variation can be accepted in a viable, reproductively successful population and that these variations, when observed in the context of geologic time, will show up as evolutionary trends. While these trends in the primates are primarily related to increasing success in utilizing the arboreal environment, they have at the same time provided essential preadaptations for ground life in some primates and man.

Primate limb development. The first group of trends relates to the development of the limbs. Although some primates show a certain amount of specialization in their degree of adaptation to a specific niche, the primates have as a rule maintained a very generalized limb and hand structure. The basic pattern of pentadactylism (five digits on each hand and foot) of the early vertebrates has been retained, as have the fibula (one of two bones in the lower leg) and the radius (one of two bones in the lower arm), bones that are reduced to vestiges or are absent in some mammalian groups. The nonspecialized primate hands and feet are mobile and prehensile; the thumb and big toe remain flexible in most groups with true opposability of the thumb possible in some. Sharp claws have been replaced by flattened nails, which improve the use of the hand as a grasping instrument. The flexible, grasping hand of the primates allows use of the forelimb for nonlocomotor activities, including important primate activities such as grooming, one-handed feeding, and infant carriage. Freeing the hands from locomotion means that the body can be held upright, and this progressive development of an erect posture is another important trend in primate evolution.

Non-specialized appendages

Primate sensory development. A further major trend in primate evolution concerns the way in which primates perceive their environment. In most land mammals the sense of smell is the one most highly developed, as is demonstrated not only in the organization of the brain but in the more obvious development of the snout. In even the most primitive primates, however, the sense of smell has lost importance, and the visual sense has become paramount. Stereoscopic vision has been achieved by moving the eyes from the sides of the head to the front of the face, thereby allowing the visual fields to overlap and producing the ability to perceive in depth. Anatomically, these changes are easy to trace in the fossil record. The importance of the eyes is demonstrated by the additional protection of a bony bar or a bony partition at the side of the head, resulting in a completely enclosed bony socket (the orbit) for the eye. This complete enclosure is absent in the closely related insectivores and in some of the most primitive living primates. The loss of importance of the olfactory sense and hence the diminution of the nose is demonstrated in the increasing orthognathism (straightness of the face) seen in the evolutionary line leading to man.

Primate dentition. Dentition has remained nonspecialized, and the occlusal (closing or biting) pattern of the molars is a simple arrangement of four or five small pointed projections (cusps). Primate dentition is composed of four distinct types of teeth: incisors (the front teeth), canines (the eyeteeth), premolars, and molars (the chewing teeth). While some groups may show specializations of certain tooth groups (an incisor "comb" in the lemurs and enlarged canines in the baboons, for example), no group demonstrates pronounced specializations involving the entire dentition as is commonly seen in other groups of mammals.

Evolutionary trends of primate behaviour. Evolutionary trends involving behaviour are obviously difficult if not impossible to document, but observations on living primate groups suggest that certain behavioral responses are more appropriate than others to the primate way of life. From such observations workers have extrapolated that certain behavioral patterns have arisen during the course of primate evolution. Virtually all living primates demonstrate a social hierarchical structure that enables them to live together in social groups of variable composition. This social structure allows—indeed demands—prolongation of infant dependency and thereby intensifies the mother–offspring relationship. It is within the primate social structure that the roots of human behaviour are to be found.

Underlying and reinforcing all of these trends is the crucial development, both qualitatively and quantitatively, of the brain. The size of the brain has, during primate evolution, increased in absolute size and in size relative to body weight. That it has also increased in complexity is demonstrated by the deeper and more varied convolutions on its surface. Those areas of the brain concerned with vision, muscular coordination, memory, learning, and communication have especially shown development.

The primate brain

As varied as these trends may seem, none developed in isolation from the others. They form a tightly interlocking feedback system whereby advances in one area are reinforced by, and in turn necessitate changes in, another area. For example, in the development of a successful adaptation to an arboreal environment, the ability to see well and to see in depth was obviously more crucial than the ability to smell acutely. The grasping hand with flat nails instead of claws is an important advantage when moving rapidly through the trees. A superior brain, stereoscopic vision, and a supremely flexible hand, along with a tendency for upright posture, were the important preadaptations (adaptations to one set of conditions that later prove to be useful or helpful under different conditions) that the ancestors of man carried with them when they left the trees for the savanna.

FOSSIL PRIMATES

Paleontology, or the study of fossils, provides the really crucial evidence concerning the evolution of the Hominidae in the past. However extensive and compelling it may be, the evidence for evolution based on the study of creatures living today can be only indirect. Further evidence of evolutionary change can be seen in the persistence of certain anatomic structures such as the vermiform appendix, which remains only as a vestige and appears to have reduced functional significance. Direct evidence of evolution must depend on actual demonstration from the fossil record of a succession of stages representing the transformation of an ancestral into a descendant type. The comparative anatomy of living forms, together with the geographic distribution of local species and varieties that exist today, suggests that evolution might have occurred. Study of the process of natural selection, experimental genetics, population statistics, and so forth establishes quite clearly how evolution could have occurred. But that evolution did occur can be scientifically established only by fossilized representative samples of those intermediate types that have been postulated on the basis of indirect evidence. Broadly speaking, the main features of the evolutionary succession of the primates are now known from the fossil record, and they conform in a remarkable way with inferences already reached by a consideration of living primates.

Primate development in the early Tertiary period. At the beginning of the geologic phase now called the Tertiary period—about 66.4 million years ago—there were in existence the most primitive of the primates. So primitive are their anatomic characters as shown by fossils that it might be impossible to determine that they were primates but for the fact that they mark a gradation toward more highly organized creatures that definitely come within the category of primates.

The climate of the early Tertiary period was warm, with wide tropical and subtropical zones extending from the equator up to the higher latitudes in both the Old and the New World. During the Paleocene epoch, which lasted for about 8.6 million years (*c.* 66.4 million–*c.* 57.8 million years ago), there were many primates in existence; 60 genera have been recognized and grouped into eight families. Three of these families had long chisel-shaped teeth that resembled those of the rodents with which they competed for a similar ecological niche, or habitat. It is likely that this confrontation was won by the rodents, for all of the early primate families with rodentlike teeth became

extinct. It is possible that this early rodent competition was one of the factors that led the primates to occupy an arboreal habitat, but a number of groups of primates, such as the baboons, the great apes, and the hominids, later returned to the ground during the evolution of the primate order.

Early Tertiary primates

Early in the Tertiary period, during the Paleocene and Eocene epochs (from about 66.4 million to about 36.6 million years ago), more advanced primates appeared that belong to the same zoological groups as the modern lemurs and tarsiers. The Lemuriformes are represented in the fossil record by a widespread family, the Adapidae, divided into two subfamilies, the Adapinae and the Notharctinae. The Tarsiiformes are known from one family, the Omomyidae, containing four subfamilies: the Anaptomorphidae, the Omomyinae, the Ekgmowechashalinae, and the Microchoerinae. It would appear that the characteristic tarsioid (tarsier-like) specialization of the skull and hind limb were already well advanced in the fossil forms that are known, but some of the European genera have some structures that indicate relationships with the early monkeys. Generally speaking, however, little is known of the Eocene ancestors of the Old World monkeys and apes; thus the Eocene epoch terminated after about 30 million years of primate evolution with lemurlike forms and tarsier-like forms but little or no evidence of anything else.

Later, in the Oligocene epoch (36.6 to 23.7 million years ago), which followed, there came into existence primitive monkeys and exceedingly primitive anthropoid apes. Excavations into the Fayum deposits of Egypt, of Oligocene age, have disclosed a deltaic shoreline bounded by tropical bush country that supported an extensive fauna including rodents, hyraxes, pigs, small elephants, and primates. One of the earliest fossil primates from the Fayum is *Parapithecus,* known from some lower jaws; *Apidium,* which is included in the same family, the Parapithecidae, is also found there. These may be the forerunners of African monkeys. Also derived from the Egyptian Oligocene are several fossil apes of primitive type; these include *Aeolopithecus,* which may be an ancestral gibbon, and *Aegyptopithecus,* which may be ancestral to the modern great apes. One other fossil ape from the Fayum that deserves special mention is *Propliopithecus,* formerly believed to be an ancestral gibbon. It has been suggested, primarily on the basis of its generalized dentition, that *Propliopithecus* is possibly ancestral to the hominids. This, however, is a speculation that requires much more evidence before it can be generally accepted.

Primate development in the Miocene epoch. The Miocene epoch began about 23.7 million years ago and lasted about 18.4 million years. It was a remarkable phase in primate evolution in which there appears to have been an increase in the numbers of larger primates that were widely spread throughout the Old World, including Europe, Asia, and Africa. The large number of specimens recovered from widely separated sites over a long period of time has led to taxonomic confusion, and more than 50 species have been described and classified in 20 genera; this number has been considerably reduced by an overdue taxonomic revision. The large Miocene hominoids appear to belong to three groups, the *Sivapithecus,* the *Dryopithecus,* and the *Proconsul* groups. Within these groups are some 24 species from Europe, Africa, and Asia.

The three Miocene hominoid groups

The *Sivapithecus* group is recognized from fossils first found in the Siwalik Hills of Pakistan and includes specimens from that region formerly known as *Ramapithecus;* other examples are known from the Çandir sites and Paşalar in Turkey, Lu-feng (Lufeng) in China, as well as from Kenya. The most remarkable specimen, with almost a complete face, comes from the Potwar Plateau of Pakistan and shows many similarities to the face of the modern orangutan. The *Dryopithecus* group includes the first specimen, found in 1856 in Saint-Gaudens, Fr. (*D. fontani*), whose molar cusp and fissure pattern, known as the Y-5 arrangement, is typical of the dryopithecines. Other examples are now known from Hungary (*Rudapithecus*), Spain (*D. laietanus*), and China (*D. keiyuanensis*).

The *Proconsul* group derives from Africa from the earlier part of the Miocene period. It includes three species—*P. africanus, P. nyanzae,* and *P. major*—as well as *Rangwapithecus gordoni* and several other smaller-bodied apes. The second group of African and Middle Miocene apes are the Oreopithecidae, which includes *Nyanzapithecus,* the large *Afropithecus,* and two species of *Kenyapithecus,* all from East Africa.

The relationships of the modern apes of Africa and East Asia to this radiation of Miocene hominoids is of prime interest, but it is complicated. It is now fairly well established that the modern orangutan is derived from Late Miocene *Sivapithecus,* based on dental and facial similarities. The date of the separation of this lineage from that which led to the African apes and to man is uncertain but seems likely to be earlier than 13 million years ago. The ancestry of the modern great apes is poorly known other than to say that the *Proconsul* group radiated early but seems too primitive in a number of respects to be linked directly with modern forms. The divergence of the modern apes may have occurred from six to 10 million years ago. The evidence obtained by molecular biologists seems to concur reasonably well on both the date of divergence of the orangutan line and that of the African great apes.

The early hominids seem certain to have originated from a Miocene ape, but it is unclear which one it might be of those that are known. It may very well be that the ancestor that is sought has not yet been discovered. The earlier view held that *Ramapithecus* was the best candidate, based on dental and palatal evidence, but such specific characteristics as thick enamel, broad, low-crowned molars, and robust jaws are shared not only with later hominids but also with several of the Middle and Late Miocene apes. (M.H.D.)

Hominidae as a zoological family

Hominidae (superfamily Hominoidea, infraorder Anthropoidea, order Primates) is the taxonomic family that includes modern humans (*Homo sapiens*) and their direct extinct ancestors. *Homo sapiens* is the only species in the genus *Homo* of the family Hominidae that is living today, and extinct populations ancestral to man are known only from fossil bones and teeth. Earlier, extinct species of the genus *Homo—H. erectus* and *H. habilis,* dating from the Pleistocene epoch (about 1.6 million to 10,000 years ago)—clearly must be included in the Hominidae; and the genus *Australopithecus*—which dates from the earlier Pliocene epoch (about 5.3 to 1.6 million years ago) and includes species such as *A. africanus, A. robustus, A. boisei,* and *A. afarensis*—is also generally included in the family.

CONTRASTING ADAPTATIONS OF HOMINIDAE AND PONGIDAE

Hominidae are distinguished from Pongidae (anthropoid apes) by evolutionary trends that illustrate the adaptations of each for different environmental situations.

Modifications of the skeleton and musculature. Obvious morphological contrasts in skeleton and musculature between living pongids and hominids suggest the functional significance of similar features observed in fossil representatives of the two families. The contrasts are particularly evident in progressive skeletal modifications in adapting to erect bipedalism. Most significant in hominids are changes in the proportions and morphological details of the pelvis, femur, and foot skeleton—all related to the mechanical requirements of erect posture and bipedal gait. Associated soft tissues also are modified, especially the hamstring and gluteal muscles and the iliopsoas muscle that flexes the trunk and thigh. The hominids preserve a well-developed pollex (thumb) and have lost opposability of the hallux (big toe).

Pongid skeletal modifications have been interpreted as deriving from hind limb reduction from a rather small ape (presumably one of the dryopithecine fossils of the Miocene epoch) whose limb proportions were those of a semi-brachiator (one that swung by the arms part of the time). The pongid pollex is reduced, and there is a strong and opposable hallux. The forelimb and hand of some pongid genera seem to be adaptations to knuckle or fist walking. Increase in size and mass of some individuals

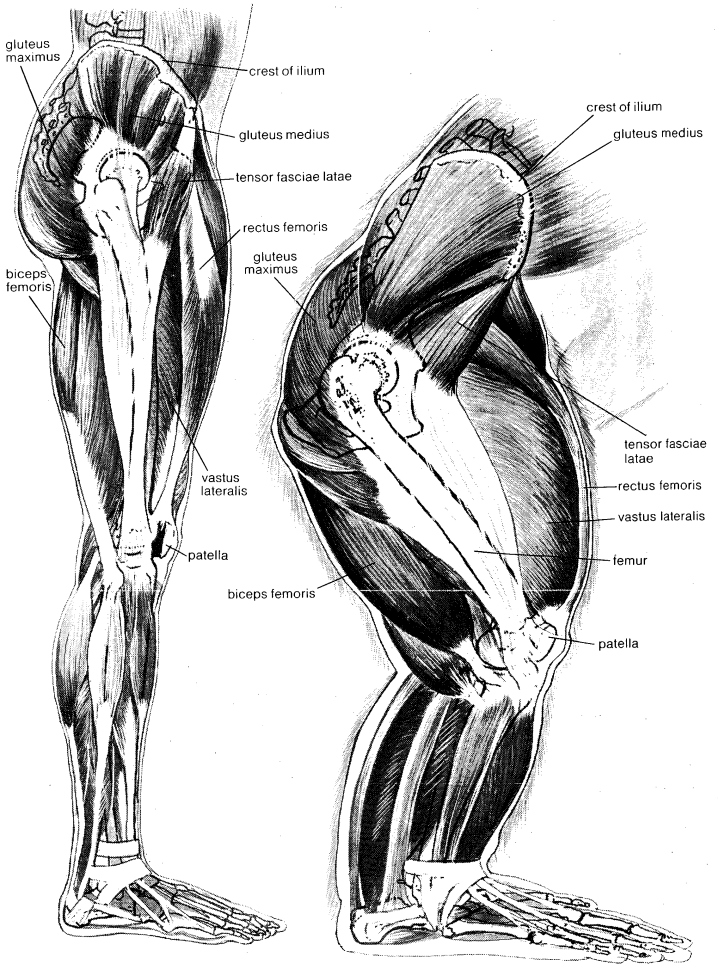


Figure 1: Right leg of (left) man and (right) gorilla.
From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

in the pongid lineage (including the large, extinct genus *Gigantopithecus*) probably made a shift from arboreal or terrestrial quadrupedal locomotion to the knuckle-walking gait a distinct advantage.

The pongid hand is still normally involved in locomotion, although in exceptional circumstances a few bipedal steps can be taken. In no pongid, however, does bipedalism form a significant part of its locomotor repertoire. In young pongids, acrobatic climbing and a form of brachiation (arm swinging) are common, but, with increase in size and weight, the African apes spend more and more time on the ground, using a pronograde (horizontal) quadrupedal stance and a gait that includes taking the weight of the upper part of the body on the knuckles or the closed fist. This specialized form of locomotion leaves a recognizable mark on the skeleton of the hand and the elbow.

The pongid pelvis retains the main proportions characteristic of all quadrupedal mammals. Although to some scholars the anatomy of the pectoral girdle (bones supporting the forelimbs) of pongids suggests a brachiating ancestor, all Miocene Hominoidea (Hominidae and Pongidae) recovered and described have failed to show specific adaptations for brachiation. At least, the Miocene hominoids do not resemble modern gibbons (*Hylobates*), whose arboreal acrobatics led to the definition of brachiation. Detailed analyses of both living and extinct pongid forms strongly suggest that the basic hominoid adaptation is dominated by forelimb locomotor and feeding behaviour, knuckle walking distinguishing pongids from their Oligocene or Early Miocene ancestors.

Skull and dentition. The hominid skull bears marks of a way of life that contrasts strongly with that of pongids. In fossil skulls of australopithecines the occipital condyles (protuberances on the base of the skull that articulate with the vertebral column) are more anterior than in pongids. Their position is associated with increased flexion of the basicranial axis, leading to upward displacement of the braincase relative to the face, with resultant increase in cranial height. These changes often are cited as serving to maintain the balance of the head, consequent to development of erect posture and bipedal gait. They are also a result of the expansion of the braincase.

The pongid skull is that of a quadrupedal pronograde primate. There is marked prognathism and, in larger species, massive jaws associated with strong muscular ridges on the skull. The extensive nuchal (nape) area of the occiput, the relatively high inion (occipital protuberance), the position of the occipital condyles well behind the level of the auditory apertures, and the limited degree of flexion of the basicranial axis are all features consistent with pongid posture. These characteristics of the skull are found in Miocene Pongidae as well.

Hominid teeth

Hominid and pongid dentitions differ significantly in form because they are adapted to different diets. Important hominid features include the reduction of the incisors and canines, the appearance of bicuspid premolars, and changes in the occlusal relationship of the jaws. The canines have diminished to a spatulate form and interlock slightly or not at all. There is no pronounced sexual dimorphism of the canines, and the spaces between these teeth (diastemata) largely have vanished. First premolars, in pongids adapted for cutting, are bicuspid in hominids, with secondary reduction of the inner (lingual) cusp in later hominid forms. Occlusal alterations tend to promote wear in all teeth to a flat, even surface at an early stage of attrition and lead to the later development of a helicoidal plane of wear. The dental arcade is even and rounded, and there is a marked tendency in later stages of the fossil record, as in modern humans, toward a reduction in molar size. Deciduous teeth appear to be replaced earlier, relative to eruption of permanent molars, and there is progressive molarization of the first deciduous premolar. This is the dentition of a terrestrial animal whose forelimbs were not primarily used in locomotion and whose dietary adaptations were for eating seeds, fruits, grasses, and meat obtained in an open savanna environment.

Pongid dentition is clearly established in the fossil record of the Miocene epoch and probably originated earlier. There appears to be progressive increase in the size of the incisors and widening of the symphysial region of the mandible with eventual formation of the simian shelf (a distinctive bony buttress of the anthropoid ape mandible). Strong, conical, sexually dimorphic canines that interlock are found throughout the pongids. The cutting function of the first lower premolar is accentuated by the development of a strong anterior root. Postcanine teeth preserve parallel or slightly divergent alignment in relatively straight rows, as opposed to the rounded hominid dental arcade. First deciduous molars remain predominantly unicuspid, and there is no apparent acceleration in the eruption of permanent canines. Pongid dentition belongs to an animal that feeds on large stalks of vegetation and fruit—tearing with its incisors, crushing with its large molar teeth, and chiseling with its enormous canines.

Skeletal, especially cranial, comparisons of living and fossil Hominidae and Pongidae once led to widely varying conclusions. Relatively few cranial characteristics of the fossils provide evidence for obviously differing evolutionary directions for the two families. For example, the cranial capacity of *Australopithecus* is relatively small, ranging from about 410 cubic centimetres (25 cubic inches) to more than 600 cubic centimetres; some skulls have strongly built supraorbital tori; and among larger early hominids a low sagittal crest occurs at the vertex of the skull (the frontoparietal area). When hominid adaptation became better understood, these features were no longer held to indicate necessarily pongid affinities. Yet it is understandable that early in the 20th-century some authorities were reluctant to classify australopithecines as Hominidae. As long as notions of what constitutes an early hominid prompt a search for a miniature version of *Homo sapiens,* most fossil hominids will be rejected. Nevertheless, hominid adaptations are clearly manifest in *Australopithecus.*

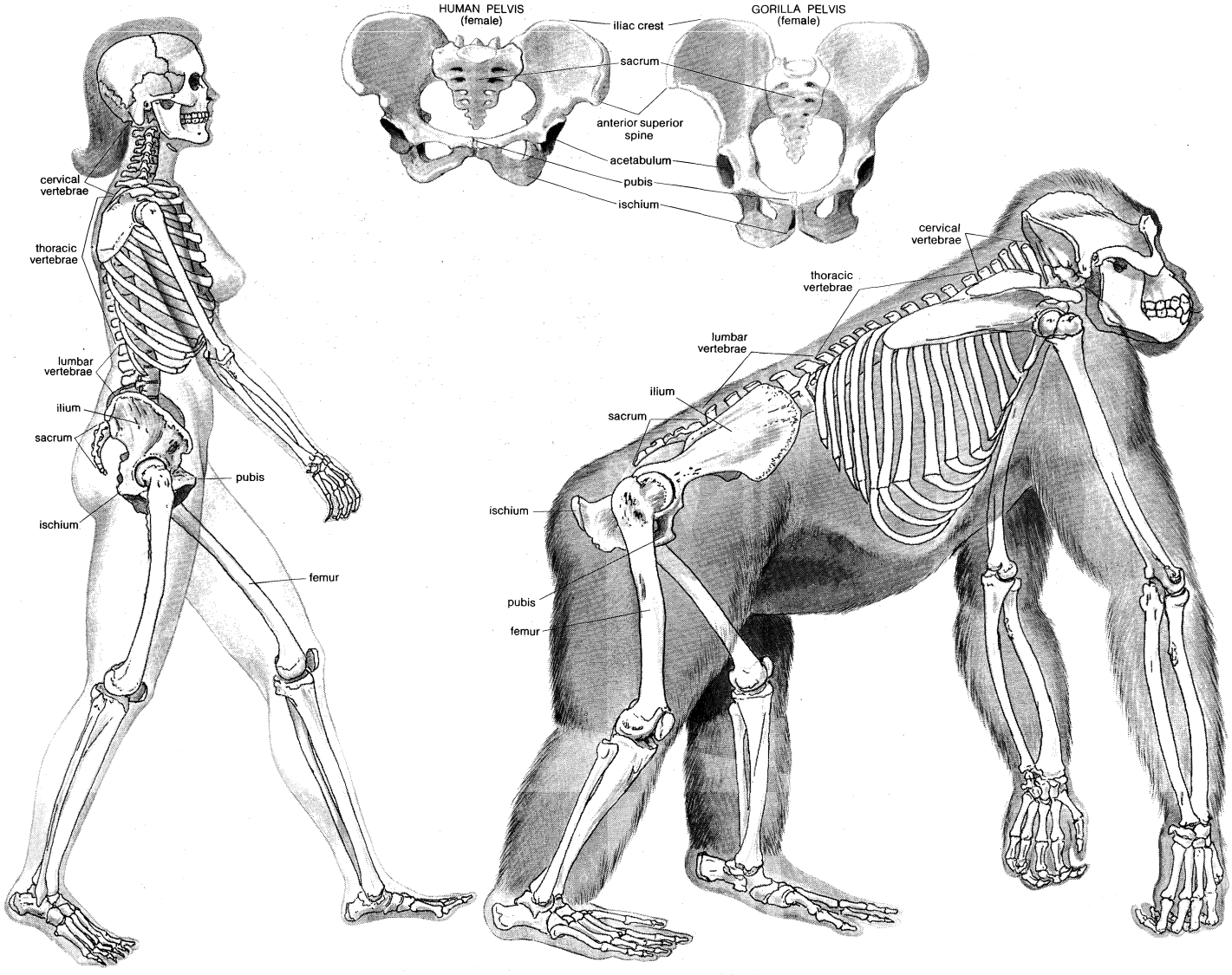


Figure 2: Locomotion of human and gorilla, relative to skeletal structures.
From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

PHYSIOLOGICAL AND BIOLOGICAL CHARACTERISTICS

The locomotor system of hominids is adapted for erect posture, bipedal gait, and manual prehensility. The hand is a highly specialized appendage adapted for a variety of grips, including a power grip for strength of grasp and a precision grip for fine manipulation. The feet are the base, in both standing and walking, through which weight and propulsive efforts are transmitted to the ground. The foot and the rest of the lower limb together form a system of levers that enables a variety of activities such as walking, running, jumping, and climbing.

Upright posture. Erection of the trunk is part of the locomotor repertoire of all primates, although many are normally quadrupedal in their resting postures and gaits. Humans and their hominid ancestors are the only primates who are, or were, habitual erect bipeds in whom the vertebral column is normally held vertically in both stance and gait.

By the late 1980s the earliest known fossil evidence of hominids came from the Middle Awash region of Ethiopia, and dated from about four million years ago. These primitive australopithecines, *Australopithecus afarensis*, were essentially erect and apparently capable of bipedal locomotion. A specimen (AL 288-1) from the Hadar site, just to the north, commonly called "Lucy," was the most complete early hominid skeleton yet recovered and showed postcranial features that permitted reliable judgments about early hominid locomotion to be made.

The Lucy skeleton and Laetoli footprints

Bipedal locomotion. The earliest known evidence of bipedal locomotion is not that of fossil hominid remains but of several trails of bipedal footprints found remarkably preserved in consolidated volcanic ash at the Laetoli site in northern Tanzania and dated to 3.6 million years ago by radiometric analysis. These footprints, when analyzed by a contouring method, disclosed that the mechanism of weight and force transference through the early hominid

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

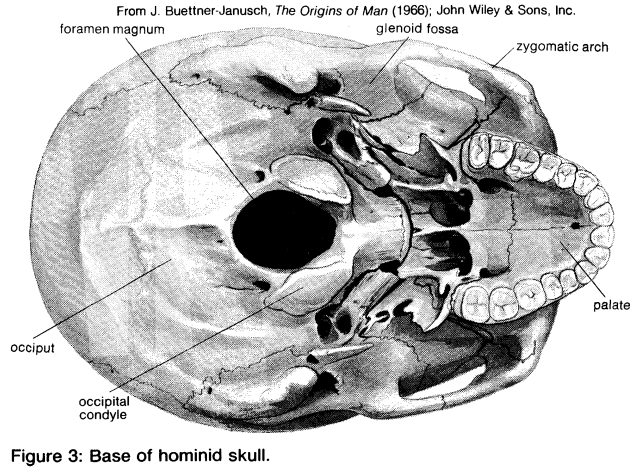


Figure 3: Base of hominid skull.

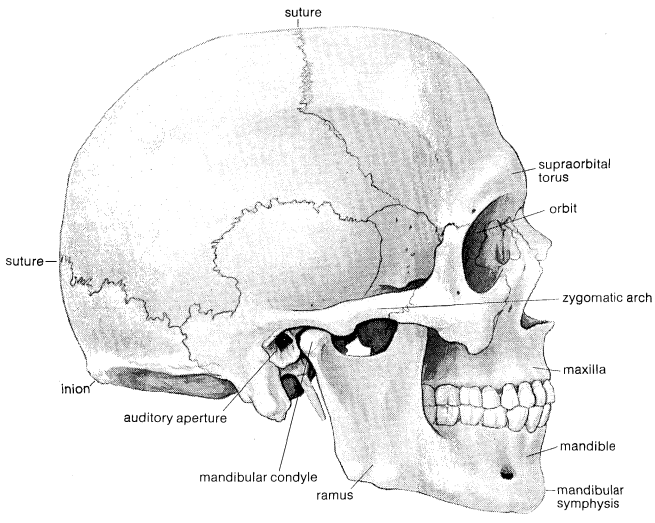


Figure 4: Right side of hominid (human) skull.
From J. Buettner-Janusch, *The Origins of Man* (1966), John Wiley & Sons, Inc.

foot was virtually identical to that of modern humans (see Figure 6). This indirect evidence of the existence and locomotor pattern of the hominids at this early date can give no indication of which group of hominids was responsible for the bipedal footprints, but other fossil evidence from Laetoli and Hadar indicates that they were made by australopithecines of an early type.

Hominid locomotor adaptations include modifications of the vertebral column, pelvis, femur, knee, ankle, and foot (for a detailed description of these, see SUPPORTIVE AND CONNECTIVE TISSUES). All of these are related to the mechanical requirements of bipedal locomotion. The vertebral column, when erect as in a biped, acquires two new secondary curves—one the cervical curve that brings the head and eyes into position for forward vision, the other a lumbar curve that permits the pelvis and lower limbs to remain in position while the trunk is erected. The lumbar curve is produced by the "wedging" of the vertebral bodies and discs, as in modern humans. This wedging is also known in australopithecine vertebrae from the Sterkfontein and Swartkrans sites in South Africa.

The hominid pelvis. The form of the hominid pelvis is regarded as a functional compromise between the needs of efficient upright stance and bipedal gait and the imperative of a female pelvis large enough to transmit a large-brained, full-term fetus.

Australopithecus and *Homo* differences

Australopithecine pelvic and limb bones differ from those of *Homo* in several anatomic features, including forward prolongation in the region of the anterior superior spine of the ilium and a relatively small sacroiliac surface. The australopithecine ischial tuberosity is usually low, and there is a marked forward prolongation of the intercondylar notch of the femur (see Figure 10).

Analysis of australopithecine and modern human pelvises allows the postulation of evolutionary changes that led to bipedalism: the ilium became shortened and bent back on the ischium; the angle between ilium and ischium grew smaller. Such bony changes seem to have been part of a process that brought the gluteus maximus muscle behind the hip joint and made it a powerful extensor of the leg.

Efforts to trace the evolution of gait and posture among the Hominidae face the critical question of whether the australopithecine pelvis is that of a hominid or more like that of a pongid. At least one characteristic of australopithecine pelvic remains is clearly hominid: the ilium is bent back, shortened, and broadened so that the iliac blade has the characteristics expected of a pelvis that is part of the bipedal complex. Investigators who favour the view that the earliest known hominid pelvis is that of the australopithecines are well aware that it is not exactly like that of the contemporary hominid pelvis in other respects. It is the major functional change, however, that assumes significance in attempts to reconstruct the past.

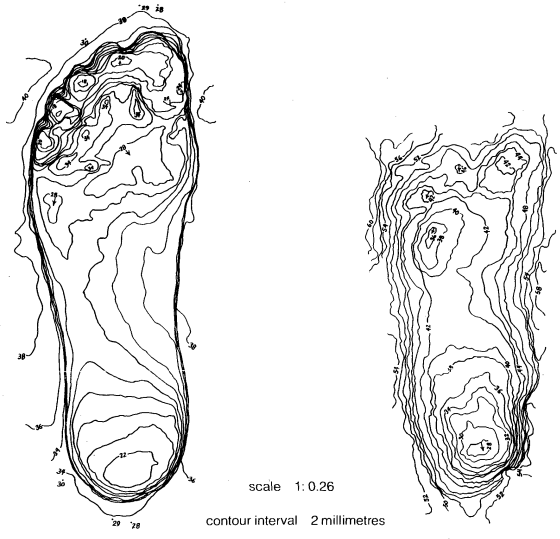


Figure 6: Comparative photogrammetric contours of (left) the left footprint of a modern human female and (right) the left footprint of a fossil hominid from Laetoli (Site G).

The femurs of modern human bipeds are obliquely placed within the thigh because of the width of the pelvis and the need for the knees to be close together during walking. This reduces the width of the walking base and avoids an inefficient, broad-based rolling gait. This requirement produces angulated femurs and knees that can be matched many times from the hominid fossil record from a variety of African sites such as Hadar and Sterkfontein, as well as Koobi Fora and West Turkana in Kenya.

The plantigrade foot

The hominid foot. Essential to the locomotor adaptation of the Hominidae is the plantigrade foot—one in which both sole and heel touch the ground—that was produced by structural modifications of the ancestral prehensile primate foot. The foot of erect bipedal man must completely support the body and be strong enough to lift the load by its lever action. Specializations that make these effects possible include the shape of the arch and the position and robustness of the big toe. Bones associated with the lever action of the foot are in such proportion that they provide support adequate to man's unique walk; humans stride. In striding, the repeated sequence in which the foot bears the weight is heel, lateral edge of the foot, ball of foot, big toe. The big toe transmits both weight and propulsive force at the end of each stride; thus the metatarsal and phalangeal bones of man's big toe are more robust than are those of the other toes.

A foot skeleton has been reconstructed from a dozen bones recovered from deposits at Olduvai Gorge that are about 1.7 million years old. While many of the proportions

From Sir Wilfred E. Le Gros Clark, *Man-Apes or Ape-Men?: The Story of Discoveries in Africa*, copyright © 1967 by Holt, Rinehart and Winston, Inc.; reprinted by permission

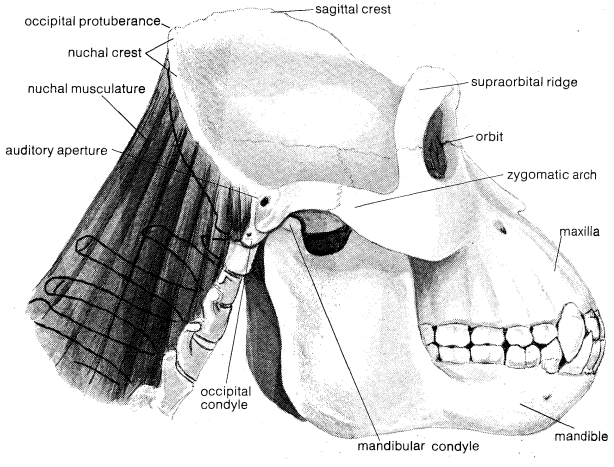


Figure 5: Right side of gorilla skull, with the massive neck muscles typical of pongids.

and articulations among the bones are typically hominid, certain bones—the first metatarsal and the skeleton of the big toe—of the Olduvai australopithecine resemble those of modern humans. The terminal phalanx of the big toe in man is specially adapted for the role that it plays in locomotion; it is unique in its form among the higher primates and contrasts strongly with that of the great apes. In line with the robustness of the first metatarsal, the skeleton of the big toe, in particular the terminal phalanx, is very stout. The terminal phalanx of modern man is broad, is flattened from above downward, and bears large collateral tubercles for the attachment of strong ligaments that support the interphalangeal joint. The bone also bears a corona that supports the large, flat big toenail that in turn supports the pad of the big toe. In addition, the terminal phalanx tilts to the outside of the foot and displays a twist in its length, both of which features are related to its function in bipedalism. The Olduvai toe bone faithfully duplicates these features and provides strong comparative anatomic evidence for bipedalism in the early hominids from Olduvai.

Forelimb structure and manipulation. The human hand is an impressive organ that distinguishes man from all other living primates, which rely on their forelimbs and hands as major organs of locomotion. Humans alone have freed their hands for manipulation by specializing for erect posture and bipedal locomotion.

The hominid hand

Fossil evidence for the evolution of the hands of hominids other than modern man is meagre. A reconstructed hand from Olduvai Gorge appears to have more similarity to those of juvenile pongids and of adult *Homo sapiens* than to the hands of adult great apes. An important feature of this hand is a truly opposable thumb—one that rotates at the carpometacarpal joint so that it opposes the other four fingers. Other significant features of this hand are the stoutness of the phalanges and their muscle markings, which indicate the presence of powerful flexor muscles; a broad, flat terminal phalanx to the thumb, which indicates a broad thumbnail supporting a substantial thumb pad; and a strong muscle marking for a long flexor tendon on the thumb terminal phalanx. This combination of features is good anatomic evidence of the strength of the grip and also of the type of grip that this creature could perform. The two main grips that are recognized in man are the power grip and the precision grip. The former is used when grasping something, such as the handle of a hammer, the latter when using a fine instrument, such as a pen or a pair of forceps.

The fossil hand from Olduvai meets the anatomic requirements of the power grip and those of the precision grip in terms of the opposability of the thumb. On this basis, it has been suggested that the Olduvai stone tools found in abundance at the Olduvai site could have been fashioned by the owner of the Olduvai hand. Hand bones recovered from the Hadar site in Ethiopia and attributed to *Australopithecus afarensis* date to about three million years ago (twice as old as the Olduvai hand) and show some signs of what may be arboreal adaption. The digits appear to be somewhat curved and thus adapted for climbing. It is uncertain, however, whether these features are signs of an active tree-living life for *A. afarensis* or of a vestigial feature from a former arboreal phase of hominid evolution.

Dentition and diet. The advent of hominid dental features in the form of the reduction of the anterior teeth (canines and incisors) seems likely to indicate a dietary shift of real significance. The amount of wear and tear on fossil teeth has always provided a clue to diet, but work with scanning electron microscopes has revealed tiny pits and scratches on the surfaces of tooth enamel—called microwear patterns—that provide further evidence of dietary conditions. It has been observed that grasses leave linear scratches on the teeth, leaves produce a polished effect, and the bone crunching of carnivores gouges out tiny pits in the enamel.

Microwear patterns on the teeth of early australopithecines seem to indicate that their diet did not consist of the tough plant material consumed by some pongids, but that the australopithecine diet had more in common with that of modern fruit-eating forms such as chimpanzees. The robust australopithecines show dental specializations of a high order in that there is gross disproportion between anterior and posterior dentition. The incisors and canines are small, while the premolars and molars are extremely large. The specialization of these molar and premolar teeth is for crushing; at first it was thought that this was an indication of increasing dietary specialization from the early to the later robust forms such as *Australopithecus boisei,* which became extinct about one million years ago. The discovery of a robust australopithecine skull from West Turkana, dated at about 2.5 million years ago, has changed that viewpoint, however, since it has the largest australopithecine molar and premolar dentition found so far as well as huge crests on the skull to permit the attachment of very large masticatory muscles. The dietary specialization of the large australopithecines, therefore, spans at least 1.5 million years.

The teeth of the early members of the genus *Homo,* such as *Homo habilis,* show changes related to both shape and proportion. The incisors become more spatulate and the molars smaller in both size and proportion, perhaps indicating another dietary shift. The teeth of *Homo erectus* show heavy wear patterns that include pits, scratches, and polish, all of which indicate an unspecialized, omnivorous dietary preference. Fossil food remains are usually the remains from meat eating, since fossil food bones are preserved more frequently than vegetable matter.

Brain and nervous system. Because brain tissue does not fossilize, the modern human brain is the only hominid brain that can be known in any detail. From the remains of fossil hominid skulls from Africa, Indonesia, China, and Europe, however, the shape and size of the hominid brain can be determined for some groups. It is clear that during hominid evolution there was an increase in brain volume and a likely increase in the ratio of brain weight to body weight. In particular, the cerebral cortex expanded in size to become a highly complex organ for the receipt, integration, and discrimination of sensory information, as well as for the initiation and coordination of motor activity. For extinct hominids, the interior of the braincase is often all that is available for studies of the volume and shape of the brain and, despite valiant efforts, the amount of reliable functional information that arises from these studies is small.

The hominid brain

Traits such as the possession of speech, the level of manipulative skills, and the level of visual and auditory abilities of early hominids cannot be judged with certainty, although some generalizations from modern man may be justified. The complexity of human behaviour is related to the human ability to interpret symbols, to appreciate abstract ideas, and to communicate them to others, particularly the young. Neuroanatomical studies show that these abilities reside primarily in the cortex of the brain, an area that has expanded rapidly in hominid evolution. Yet volume alone is not enough, and the level of differentiation and organization of brain tissue may also be of crucial importance. The fossil record can yield only endocranial casts and, from them, possible brain volumes, but the firm association of stone tools with such remains must indicate a level of intellectual attainment that can foresee a use for a tool, envisage it within a stone, and then shape it to a set and repeatable pattern.

Biomolecular characteristics. Chromosomes, hemoglobins, blood groups, many serum proteins, and red-cell enzymes, among other genetically controlled traits, have been studied extensively in contemporary *Homo sapiens* and the modern pongids. No reliable information on homologous traits of any extinct hominid exists, since chromosomes and proteins do not fossilize. Data available for man and living pongids, however, are compatible with modern notions of the close affinity between hominids and pongids. Precise immunologic and biochemical comparisons, representing efforts to refine views of the phylogenetic relationship of pongids and hominids, have been inconclusive at best. Use of these data to specify the time of divergence of the hominids and the pongids has produced dates at variance with paleontological opinion based on the fossil record. The theory of late divergence—from five

to four million years ago—favoured by evolutionary biochemists, however, has become more favourably viewed by some paleontologists than it was in the past.

BEHAVIORAL CHARACTERISTICS OF HOMINIDAE

Homo sapiens is biologically close to the other higher primates, but it is behavioral characteristics that predominantly set the species apart. The most significant features of hominid behavioral evolution are the development of toolmaking capabilities, conceptual thought, and symbolic language. These features are relatively simple to examine in *Homo sapiens,* but in extinct hominids they can only be inferred indirectly from fossil remains.

Toolmaking capabilities. The ability to use tools is not unique to the Hominidae; vultures are reported to use rocks to crack open ostrich eggs, and sea otters use stones to open clam or oyster shells. It was long claimed that only hominids actually made tools, but in 1960 chimpanzees were observed in the wild breaking off twigs from trees, stripping away the leaves, and using the twigs to extract termites from their nests. Since the twig had to be modified by removing its leaves, this activity constituted toolmaking, even though at a crude level. Repeated observations of such behaviour resulted in hominids being defined not only as toolmakers but also as skilled toolmakers; they are the only animals able to use one tool to make another and to manufacture standardized precision tools.

Tool use obviously preceded toolmaking in hominid development, but—because the earliest tools were probably made from perishable materials, such as wood—it is not known when either of these activities became a regular feature of hominid behaviour. The first recognizable tools of hominids were made of stone, which, fortunately for archaeologists, is extremely durable. When a fine-grained stone without cleavage planes is hit with another, it will always break in a distinctive fashion, called a conchoidal fracture. Characteristic of this fracture pattern are a bruised striking platform at the point of impact with shock waves radiating from it and, on the resultant flake, a bulb of percussion and bulbar scar. When these features are present, it is possible to distinguish human workmanship from natural breakage caused by heat or frost. Stone tools first appear in the archaeological record as crude pebble choppers with no more than a few flakes removed from one side; they have been found at sites with fossils that date to at least two million years ago. By the Middle Pleistocene the tools had developed into finely made, symmetrical, ovate hand axes that were flaked on both sides and carefully trimmed or retouched to produce straight edges. These are the first standardized artifacts (*i.e.,* objects made systematically to conform to a pattern preconceived in the maker's mind) and are the first to show that the toolmakers' considerations were aesthetic as well as functional. With the development of the soft-hammer technique and the use of pressure as well as percussion flaking, stone tool technology continued to be refined, and hafting was introduced to produce composite tools with wooden handles. The delicate blades and arrowheads of the Upper Paleolithic (*c.* 40,000 to *c.* 10,000 years ago) represent the pinnacle of stone technology, when functional considerations were sometimes overshadowed by aesthetic or symbolic values. Modern man has, of course, superseded stone tools with those made of metals.

Toolmaking from stone

No tools have been found in association with the hominid fossils from Laetoli and Hadar attributed to *Australopithecus afarensis.* The earliest tools, from sites such as Koobi Fora and Olduvai Gorge, have been found in stratigraphic levels that have yielded robust australopithecines, *Homo habilis,* and—in the case of Koobi Fora—*Homo erectus.* Under these circumstances it is not possible to decide which of the hominid species was the first stone toolmaker.

Language and symbolic behaviour. Toolmaking at least leaves behind some tangible evidence, but it is much more difficult to trace the origins of spoken language in the Hominidae. Inferences about the anatomy and position of the larynx can be made from cranial form, but otherwise fossil bones tell us nothing about speech capabilities. Archaeological finds of standardized tools and evidence of the cooperative hunting of large animals may well indicate some kind of communication among early hominids, but not necessarily in the form of modern human language. All primates, including man, use visual communications such as facial expressions, body language, and nonlinguistic vocalizations such as screams and cries to transmit information to other members of their group. These communications are largely instinctive, and the signs are very limited in meaning. Spoken language, however, allows human beings to name things with "open" symbols—*i.e.,* symbols that, in countless combinations, can be made to relay different messages. Not only can human communication cover immediate situations and feelings but also discussions at abstract or hypothetical levels. Humans thus can store and transmit knowledge gained by past experiences as well as discuss plans for the future.

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

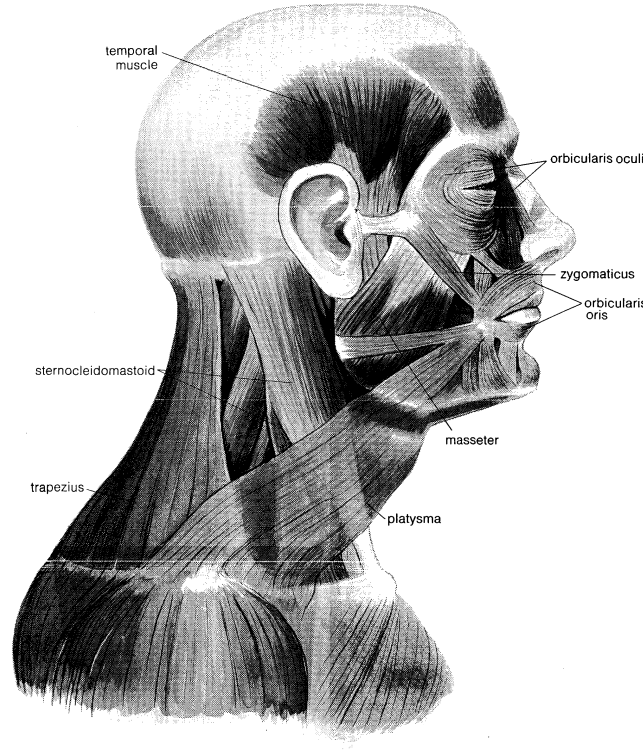


Figure 7: Muscles of human head and neck.

The principal difference in the vocal apparatus of hominids and pongids is the position of the larynx relative to the rest of the respiratory tract. Man's larynx is lower in the throat and farther from the soft palate than in other primates, directly reflecting development of erect posture and expansion of the braincase. The position of the larynx changed as the foramen magnum (the skull's aperture for the spinal cord), in turn, moved in response to the way the head is balanced on the neck and to the expansion of the posterior portion of the base of the skull. The consequent descent of the larynx created a long, tubular resonating cavity that permits the low-pitched speech of man.

Possible origin of hominid language

While no evidence exists to show how hominid language first developed, it has been suggested that the vowel sounds had their origins in nonlinguistic vocalizations and that consonants were added as the hominids developed more control over their airways by manipulating tongues, lips, and teeth. Coupled with the ability to make the sounds necessary for speech would be changes in the brain that allow vocabulary to be stored and retrieved and changes in the auditory apparatus that allow language to be understood when spoken by others with slightly different intonation or pitch. The ability to learn language is inherited, as all children learn very quickly, but the actual language, be it English or Chinese or any other, must be learned. Written language is clearly a much more recent human acquisition, with no good evidence for it before 5,500 years ago.

Because the fossils reveal so little, studies of the develop-

ment of hominid speech have centred around experiments with one of man's closest living relatives, the chimpanzee. The structure of the chimpanzee larynx does not allow it to make many of the sounds needed for human speech, and so experimenters have concentrated on teaching chimpanzees sign languages designed for deaf people. One female chimpanzee, Washoe, was taught to make signs for more than 100 words and showed that she could understand more than 300 signs. She was eventually able to put two or three signs together to make rudimentary sentences, usually demands for food, but did not make signs spontaneously in order to communicate ideas.

Social organization. The shift to terrestrial life in an open savanna environment would mean that a small, upright biped without the benefit of large, powerful jaws would be vulnerable to attack from predators. Opposing views have been proposed to explain why the early hominids not only survived but also prospered. One view is the "hunting hypothesis," which suggests that keen eyesight, cunning, and the development of weapons necessary to hunt large animals channeled the natural aggression of humans and enabled them to defend themselves against predators. The other view stresses the significance of hunting as a stimulus toward social cooperation and food sharing.

Also important to the second view is the fact that human infants are much more helpless than other primate babies and require constant care for a relatively long period of time. With the formation of kinship groups a division of labour could occur, with the men doing the hunting and the women staying behind to gather plant foods and care for the children. Pair-bonding within groups may have been established among the early hominids when the females became sexually receptive at all times, unlike other primate females who have an estrus cycle. Members of a group who perform different economic tasks during the day must operate within a limited territory and have a place where the entire group reassembles to distribute the food and spend the night. The concept of a home base has been confirmed in the archaeological record from early sites at Koobi Fora, where concentrations of stones and bones indicate places that were continuously or periodically occupied by groups for the purposes of butchering and eating animals and making tools. (J.B.-J./M.H.D.)

THE EVOLUTION OF HOMINIDAE

Australopithecus

Australopithecus (literally "Southern Ape") was the generic name given to the first-discovered member of a series of fossils of creatures closely related, if not ancestral, to modern human beings. Since the first discovery—of a child's skull in a cave at Taung, S.Af., in 1924—similar hominid remains have been found at numerous sites in East and southern Africa. The term australopithecine is often used to refer to all the fossil hominid material that dates between the last half of the Late Miocene epoch (about eight to 5.3 million years ago) and the beginning of the Pleistocene epoch (around 1.6 million years ago). Fossil remains that date from before eight million years ago are widely regarded as those of fossil apes, while evidence of *Homo erectus* ("Upright Man") dates to about the beginning of the Pleistocene. The remains of *Homo habilis* ("Handy Man") have been recovered from sites between 2.5 and 1.5 million years old. These are more easily confused with those of *Australopithecus,* and the final allocation of some of this material is still being debated.

The fossil evidence of the australopithecines has been

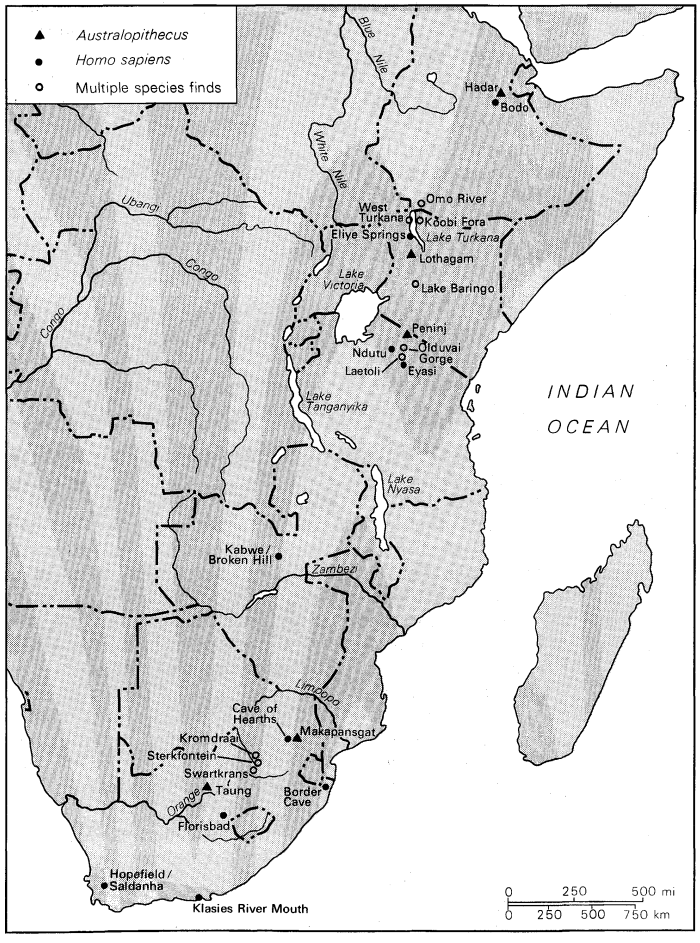


Figure 8: Major sites of hominid fossil finds in sub-Saharan Africa.

seen by some scholars as merely representing temporal stages within a single evolving hominid lineage leading to *Homo erectus* and thence to *Homo sapiens.* Others have stressed the extent of the adaptive differences between the various fossils and have suggested that there may have been two, or even three, lineages evolving in parallel, only one of which led to the later species of *Homo.* Whatever the details of their interpretations, however, most hominid paleontologists are agreed that the australopithecines represent a link—direct or indirect—between the fossil apes and human beings. Thus, the study of the australopithecines is regarded as the study of one of the most important stages in the emergence of modern *H. sapiens.*

FOSSIL EVIDENCE

The South African australopithecines. More than a decade was to elapse between the recognition of the importance of the Taung child's skull by Raymond Dart in 1924 and the next series of discoveries of australopithecines in southern Africa. These latter discoveries were made by Robert Broom in 1936 and 1938 as the direct result of mining operations at the caves of Sterkfontein and Kromdraai, several hundred miles northeast of Taung, in the Transvaal. When research activities resumed in earnest in South Africa after World War II, two additional cave sites were discovered at Swartkrans and Makapansgat. After the early discoveries the rate of recovery of fossils from these hard, breccia-filled cave deposits diminished. During the 1970s, however, there was renewed activity at these sites, and the total number of hominid remains recovered from southern African caves is well in excess of 1,000.

As each series of discoveries was announced, it was usually marked by the suggestion of a new species, if not genus, for the newly found fossils. Scientists are now agreed, however, that the evidence does not justify the multiplicity of taxa that resulted. It is generally accepted that the australopithecine fossils recovered from these cave sites belong to either *Australopithecus africanus,* the species usually referred to as the "gracile" australopithecine, or *A. robustus,* the species called the "robust" australopithecine. (Some workers support the recognition of a second species of robust australopithecine found at Swartkrans, called *A. crassidens,* but this is a minority view.) The classification into two main species within one genus won the support of Dart and Sir Wilfred Le Gros Clark, and, with minor modifications, it is the scheme that has come to be supported by the majority of scholars. The labels gracile and robust are useful because they are widely understood, but they also are dangerous in that such a simple classification is misleading. An increasing number of workers have argued that the generic name *Paranthropus* ("Next to Man"), coined by Broom in 1938, should be revived to distinguish the robust australopithecines.

Two main species

The skull of *Australopithecus africanus* has a braincase that is roughly spherical in shape, with the greatest width across the base of the skull. The cranial capacity of the species depends upon which estimates are used; the average value is probably between 430 and 450 cubic centimetres (26 and 27 cubic inches), a capacity just within the range of that of living apes. Brain size is best judged in relation to body size, however, and on this count the gracile australopithecines have a relative brain size intermediate between that of modern apes and that of modern human beings. The area of attachment on the skull for the neck muscles is reduced when compared to that of equivalent-sized apes. Although bony crests mark the attachment of the jaw muscles onto the skull, these are not obvious features and, in particular, do not form large midline, or sagittal, crests. The foramen magnum (the area of the skull through which the spinal cord passes) lies nearer the centre of the skull in these fossil hominids than it does in the apes. The face is projecting, but it does not form the marked muzzle that is a feature of most modern and fossil ape skulls.

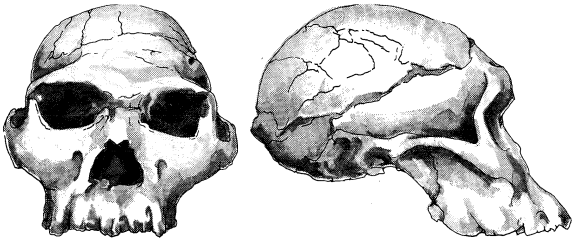


Figure 9: Skull of *Australopithecus africanus,* specimen STS 5, found at Sterkfontein, S.Af.

The teeth of the gracile australopithecines are not arranged in the characteristic U-shaped fashion of the apes but instead lie in a more rounded arcade. The incisor teeth are set vertically in the jaw, and the canines are small and do not project above the other teeth, which is always the case in the apes. There is no gap, or diastema, between the canines and the premolars, and the upper canines do not form a shearing unit with the first lower premolars. The milk teeth resemble those of the later hominids, but the order of eruption and the rate of maturation of the teeth follow the ape pattern more closely than they do that of modern humans.

There are fewer limb-bone fossils than skull remains. Upper-limb bones are particularly poorly represented and provide little or no information about manipulative ability, but there are indications from remains of the shoulder that this region of the gracile australopithecine was well adapted for climbing. The remains of the lower limbs and vertebral column provide good evidence for a more or less upright posture and suggest that these creatures walked bipedally in a way that was more efficient than the occasional bipedal walking observed in apes. An important aspect of this evidence is the low, posteriorly expanded blade of the ilium with the characteristic sciatic notch. The greatly increased width of the pelvic cavity, when compared to that of the apes, allowed for the birth of infants with larger heads and is additional evidence of the increase in relative brain size (see Figure 10). Estimates of stature and body weight are necessarily imprecise, but they suggest that the height of the gracile australopithecines was about 150 centimetres (five feet) and that they weighed between 35 and 60 kilograms (about 75 and 130 pounds).

Evidence of upright posture

The robust australopithecines found at sites in southern Africa share many of the basic features of the gracile group, the main points of difference being in the skull and the teeth and jaws. The average brain size of the robust australopithecines is a little larger than that of the gracile form and averages about 500 cubic centimetres. Because estimates of the robust form's body weight are between 10 and 25 percent greater than those of the gracile form's weight, however, the relative brain size of the two species is of the same order. The skulls of the robust australopithecines are more rugged than those of the graciles; crests that mark the attachment of the jaw and neck muscles are better developed, and the face is flatter and broader.

The dental arcade of the robust sample differs from that of the gracile remains in several ways. The molars of the robust form are larger, and the premolars of the lower jaw tend to develop extra cusps and so appear more like molar teeth. The anterior teeth of the robust form show no corresponding increase in size; in fact, their average size is smaller than in the graciles. The canines of the robust dentition are conical and more like those of later hominids, while the canines of the graciles are unusually asymmetrical. All of these features, taken with microscopic evidence of tooth wear, suggest that the face of the robust australopithecines had become specialized to increase the tooth area devoted to chewing, concentrating the power of the jaw muscles on the molar and premolar teeth.

The extent and functional significance of the differences between the pelvic and hip regions of the two types of australopithecines has been vigorously debated. John Talbot Robinson has contended that the robust australopithecine pelvis is adapted more for activities that emphasize power, such as climbing, than for length of stride and is thus significantly different from that of the gracile form. He and others have also maintained that the shape of the top end of the femur is distinctive. Many other scholars, however, interpret the fossil evidence as indicating no major differences in the gait of the two types of australopithecines, contending that there is still too little well-preserved ma-

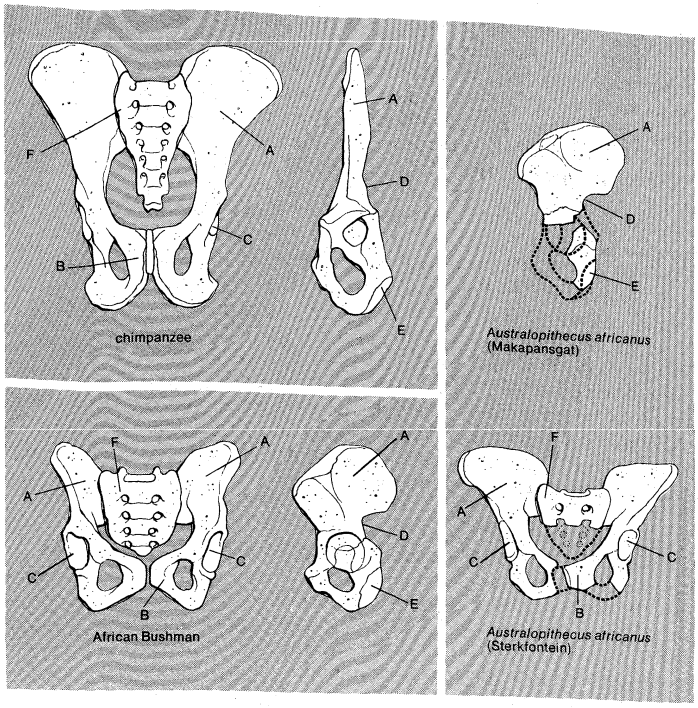


Figure 10: *Front and side views of pelvis.*
(A) Iliac blade. (B) Pubic bone. (C) Acetabulum (hip joint).
(D) Sciatic notch. (E) Ischium. (F) Sacrum.
From W.E. Le Gros Clark, *The Fossil Evidence for Human Evolution* (1964);The University of Chicago Press

terial to justify claims for substantial differences in locomotor habit.

The East African australopithecines. The first major discovery of australopithecine remains in this region was made at the Olduvai Gorge in Tanzania by Mary and L.S.B. Leakey. In 1959 they discovered a well-preserved cranium—designated Olduvai Hominid (OH) 5—that showed, in an exaggerated form, many of the features of the robust australopithecines from Swartkrans and Kromdraai. Although the cranium was initially attributed to a new genus and species—*Zinjanthropus boisei*—it was later suggested that the fossil be included as a separate species in *Australopithecus*. The discovery in 1964 of a massive robust australopithecine jaw at the Peninj site near Lake Natron in Tanzania confirmed the presence of this species in East Africa. Remains of a larger-brained, smaller-toothed hominid were also found at Olduvai from the early 1960s onward, and some have believed that these represented *A. africanus*. More recent analyses, however, have provided support for classifying the remains as *Homo habilis* (see below *Homo habilis*).

Olduvai Gorge discoveries

Since then, the major contributions to knowledge of the australopithecines have come from research at sites on the Omo River in Ethiopia; at Koobi Fora (formerly East Rudolf), on the northeastern shore of Lake Turkana (Lake Rudolf); and at West Turkana, on the northwestern shore of Lake Turkana. At the first two sites there is evidence of both gracile and robust hominids. The more robust remains clearly belong to *Australopithecus boisei,* the East African variety of the robust australopithecines. There is a good deal of size variation among the skulls, jaws, and teeth belonging to *A. boisei.* This suggests that the males and females of this taxon may have been markedly different in body size, which has implications for any attempts to reconstruct the social organization of these creatures. The remains of a nearly complete cranium from West Turkana (designated KNM-WT 17000 and popularly called the "black skull") also have been attributed to *A. boisei,* but they may exemplify a different and less-specialized taxon called *A. aethiopicus* (see below).

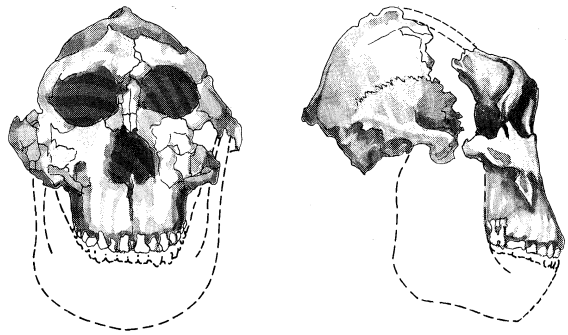


Figure 11: Skull of Olduvai Hominid (OH) 5 from Bed 1, Olduvai Gorge, Tanz.

The interpretation of the more gracile, smaller-toothed remains is more problematic. Evidence from the earliest fossil-bearing strata at Omo and Koobi Fora consists mainly of isolated teeth, and they most closely match teeth belonging to *A. africanus.* The gracile fossils from slightly younger strata at both sites are more difficult to interpret and resemble in many ways the *H. habilis* material recovered from Olduvai Gorge. Some of the cranial remains (*e.g.,* specimens KNM-ER 1470 and ER 3732 from Koobi Fora) suggest that the brains of these creatures were significantly larger than those of *A. africanus,* yet other gracile crania (*e.g.,* KNM-ER 1813) have brain capacities between 500 and 600 cubic centimetres, clearly within the accepted range for gracile australopithecines. Jaws and teeth show a mixture of australopithecine and later *Homo* features, and some of the limb bones suggest affinities with later *Homo erectus* remains. The result of this mosaic of features is that opinions differ about how this material should be classified. Some scholars regard it as strengthening the case for *H. habilis,* yet others consider it to be a geographic variant of *A. africanus.*

The Koobi Fora skulls

Fossil evidence of australopithecines from Laetoli in Tanzania and from Hadar and Middle Awash in Ethiopia represents hominids in the 4.5- to 2.5-million-year time period. In the 1930s the site at Laetoli (then called Garusi) yielded fragmentary remains of two upper jaws. The site was reexplored in the mid-1970s by Mary Leakey, resulting in a series of fossil hominid discoveries that included trails of hominid footprints (see below). Fieldwork carried out in 1972–77 by an international expedition led by Donald Johanson, Maurice Taieb, and Yves Coppens at Hadar, an extensive fossil site located in the Afar Triangle of Ethiopia, resulted in a remarkable collection of several hundred fossil hominid remains. Two discoveries were particularly notable. One is an individual specimen (AL 288-1, popularly called "Lucy") that includes nearly half of the bones of the preserved skeleton; the other is a series of fossils, popularly called the "First Family," from locality AL 333 that includes remains from at least 13 individuals.

The initial assessment of the material from Laetoli suggested that the remains most closely resembled the allegedly early *Homo* fossils from Koobi Fora, and thus the Laetoli evidence was also tentatively referred to the genus *Homo.* Similar inferences were made about part of the Hadar sample, while the remaining material from Hadar was considered to show affinities with the gracile and the robust australopithecines. More-detailed examination, however, resulted in a new interpretation that links the remains at the two sites in an entirely new australopithecine species, *Australopithecus afarensis.* The authors of this proposal, Johanson and Timothy White, listed a series of features that they consider as differentiating *A. afarensis* from *A. africanus,* the existing taxon that it most closely resembles. The features they find important include a more projecting face; a long, narrow, straight-sided dental arcade, with relatively and absolutely large anterior teeth (canines and incisors); and the shape of the canine and premolar teeth.

Proposal of *Australopithecus afarensis*

The distinctiveness of these features that, together with aspects of the mandible and cranium, make up the diagnosis of *A. afarensis* has been challenged by other researchers. In their view the features cited either overlap with the known samples of *A. africanus* or are features that are common to all australopithecines. Such shared features, these researchers argue, are of no value in defining a single species within the same group of fossil taxa. Some researchers have also pointed to the large range in size

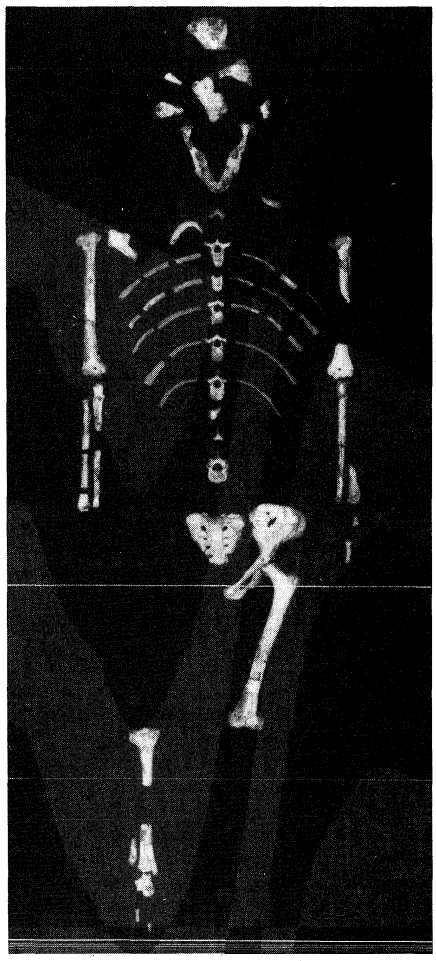


Figure 12: The AL 288-1 ("Lucy") skeleton, found at Hadar, Eth., and attributed to *Australopithecus afarensis*.
Cleveland Museum of Natural History

of mandibles and teeth in the Hadar collection and have questioned whether such variability can be accommodated within a single taxon. Subsequent research reports, however, have consistently lent support to the interpretation that the specimens from Laetoli and Hadar represent a single species that is distinct from *A. africanus*.

DATING THE FOSSILS

The most useful absolute-dating methods for australopithecine finds have been through potassium–argon and uranium-series dating. The record of the Earth's magnetism preserved in strata has also been widely used to date fossil hominid sites. The most common method of relative dating (*i.e.*, finding links between the strata of one site and those of another, well-dated locality that is used as a reference) uses fossils of ubiquitous species, such as fossil elephants and pigs, to provide relative dates for australopithecine-bearing hominid sites.

Dating the South African caves. The cave sites in South Africa have proved difficult to date by absolute methods because the sediments they contain lack most of the useful dating isotopes and the cave fillings seldom preserve a sufficiently strong or reliable record of the direction of the Earth's magnetic axis. Another problem is that the stratification within the caves is complex, with interposition of the strata. The estimated ages of the australopithecine-bearing strata at these cave sites are given in Figure 13. The oldest is probably Makapansgat, its dating being one of the few for the cave sites that combined absolute and relative methods. The vast majority of the australopithecine remains from Sterkfontein come from the stratum called Member 4 and are dated to three to 2.5 million years ago. Kromdraai is usually considered to be younger, though a slightly older date for these hominids is possible.

Dating the East African sites. Attempts to date hominid sites in East Africa have not been without problems. Nonetheless, it is widely accepted that, because the region's isotope-rich layers of volcanic ash, called tuffs, have provided good samples for absolute dating methods, the dates for East Africa are more reliable than those for the cave sites in South Africa. Each tuff has its own chemical profile, and tuff "fingerprinting" has enabled ash layers from sites hundreds of miles apart to be linked to the same eruptive event.

Uncertainties about the classification of the gracile fossils from Olduvai, Koobi Fora, and Omo mean that the attributions of taxa to sites shown in Figure 13 are necessarily simplified. Two observations are important, however. First, in both East and southern Africa the gracile australopithecines, *A. africanus* and *A. afarensis*, antedate the robust australopithecines. Second, the emergence of the robust australopithecines at the East African sites seems to coincide with the appearance of the so-called advanced gracile hominids attributed to *Homo habilis*.

ARCHAEOLOGICAL EVIDENCE

The relative richness of the archaeological record at the principal East African australopithecine sites contrasts with the lesser amount of evidence recovered from the cave sites of southern Africa. The probable explanation for this has come from careful studies of the animal bones found together with the hominid remains in the South African caves. The results of these studies suggest that the australopithecines were not living in the caves but were simply part of the bone refuse accumulated by a predator, most likely a leopard-sized creature. Even if these gracile and robust australopithecines had been making stone artifacts, it is most unlikely that their tools would have found their way into a carnivore's lair.

Bone "tools" from Makapansgat. Perhaps the best-known so-called archaeological evidence from the South African cave sites came from Makapansgat. Two discoveries there were thought to be particularly significant. The first was an unusual abundance of the jaws and forelimb bones of fossil antelopes, and the second was a series of baboon skulls that had been extensively fractured. The evidence of the antelope bones led Dart to propose the existence of what he called the osteodontokeratic ("bone-tooth-horn") culture, speculations that later were to foster the idea that the early australopithecines were "killer apes." Research on the way predators and scavengers deal with animal skeletons, however, has cast considerable doubt on this interpretation. These findings suggest that the natural breakup of antelope skeletons leads to the differential survival of particular bony parts, and that this,

The killer-ape theory

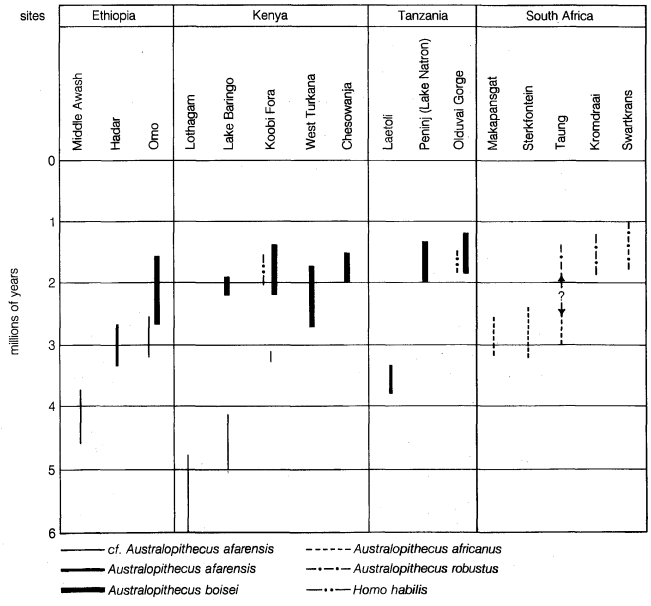


Figure 13: Approximate time ranges of sites yielding *Australopithecus* and *Homo habilis* fossils.

along with the preferences of predators and damage to the bones by stones falling from the cave roof, would be sufficient to account for the bone accumulations and the pattern of breakage found at Makapansgat. These studies do not exclude the use of bone tools at Makapansgat, but they do suggest that the present evidence for it is, at the very least, ambiguous.

Stone artifacts. No stone artifacts have been found at Makapansgat, but small collections of such artifacts have been recovered in the later robust-australopithecine-bearing layers at Swartkrans and in strata of similar age at Sterkfontein. The Swartkrans artifacts are mainly relatively crude stone chopper cores, flakes, and scrapers made of quartzite and quartz, and some bone tools. The artifacts at Sterkfontein, also made of quartzite, are technologically more advanced and include small scrapers and primitive hand axes. Subsequent analysis of the hand bones from Swartkrans—which are presumed to be australopithecine—has demonstrated that they are compatible with tool use.

Although artifacts have been recovered from older levels at Hadar, the earliest substantial and comprehensive archaeological evidence from East Africa occurs in strata dating from around two million years ago. The earliest of this series of occurrences is at the Omo River in Ethiopia, where collections of small quartz flakes have been found. It is, however, the sites at Olduvai Gorge and Koobi Fora that have provided the richest evidence of early hominid technology and behaviour. At both sites there is evidence that by about 1,750,000 years ago hominids had developed sufficient manipulative skills and cognitive ability to fashion a range of stone tools. The best known of these early industries is the Oldowan, first described from sites in Bed I of Olduvai. The artifacts range from choppers to small flakes, and variants (and perhaps developments) of this basic tool kit have since been described in later strata at Olduvai and other East African sites.

The Oldowan industry

The lack of good lithic tool evidence at early australopithecine sites and the contemporaneous appearance in the fossil record of *Homo* and stone artifacts led most workers to presume that *Homo habilis* manufactured the artifacts. There is, however, nothing other than circumstantial evidence to substantiate this association, and the most recent finds from Swartkrans suggest that the behavioral contrast between *Homo* and *Australopithecus* may be related to the different ways in which the two forms used stone tools.

HABITAT

Since the 1960s scientists have paid particular attention to reconstructing the paleoenvironment of the australopithecines. Information about the strata, fauna, and flora (from pollen analysis) of the period have been combined to provide reliable assessments of the habitat of the early hominids.

The evidence from both East and southern Africa suggests a similar pattern of climatic change. Bovids, which include the antelopes and bucks, are especially sensitive to environmental change. Their distribution in the southern cave sites suggests that the deposits in the earlier sites of Makapansgat and Sterkfontein (Member 4) accumulated under more wooded conditions than those of the later sites of Swartkrans and Sterkfontein (Member 5), where the types of bovids found indicate a drier climate with more open grassland. Evidence from the early sites in East Africa seems to indicate an open woodland environment, but the results from studies at Omo, Olduvai, and Koobi Fora suggests that a marked change in climate and vegetation occurred throughout East Africa about 2.3 million years ago. The evidence, ranging from oxygen isotope analysis of lake sediments to the detailed examination of microfauna, all points to a shift at that time to a drier climate and a more open, relatively treeless, scrub-type savanna environment.

BEHAVIORAL INFERENCES AND EVOLUTIONARY IMPLICATIONS

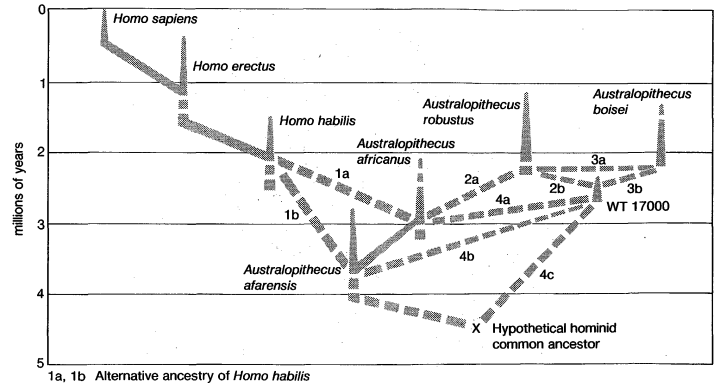
Fossil and paleoenvironmental evidence from the Miocene epoch, the finds at Hadar and West Turkana, and the bipedal footprints and other remains from Laetoli have led paleontologists to revise their theories about the emergence and adaptations of the australopithecines and of their evolutionary relationships.

Emergence and adaptation. The paleoenvironment at most of the fossil-ape-bearing Miocene sites included open woodland, similar to the habitats reconstructed for the early australopithecine sites. The footprints at Laetoli are unambiguous evidence for bipedal walking, and the pelvic remains from Hadar confirm that modifications for upright posture were well established by three million years ago. There is, however, increasing evidence from studies of the limb bones of the australopithecines that the skeletal adaptations for climbing and bipedal walking are similar. That the skeletal changes necessary for climbing may be preadaptive for bipedal running and walking suggests that the early gracile australopithecines from East and southern Africa may have used both these modes of locomotion and may well have spent more time resting and feeding in trees than has hitherto been believed. The dental remains of the gracile australopithecines do not suggest any particular dietary adaptation, and researchers have suggested that they subsisted mainly on a plant diet consisting of fruits, berries, and tubers. The most reasonable hypothesis to account for the appearance of the gracile australopithecines prior to about four million years ago is that their appearance coincided with a climatic change about five million years ago to a generally drier climate; that shift marked the beginning of the desertification of the Sahara.

Bipedal locomotion

A second climatic shift, the one noted above that is associated with further expansion of the savanna environment, coincided with the emergence of the robust australopithecines as well as with the appearance of the more advanced hominids attributed to *Homo habilis* and perhaps to other species of *Homo.* Studies of the molar teeth belonging to robust australopithecines have suggested that they would have been well adapted to crushing hard objects, such as the casings and husks of fruits. Thus, it may be that the drier climate forced the robust australopithecines to occupy a specialized—and ultimately limited—niche in the sparsely wooded savanna grasslands. During the same period it seems most likely that the dentally less specialized australopithecines were adapting in a different direction, and it is natural to link them with the quickening pace of cultural advance and the emergence of *Homo habilis* and, later, *Homo erectus.*

Evolutionary relationships. The various phylogenetic hypotheses, or trees, that have been put forward to explain the relationships between australopithecine taxa are summarized in Figure 14. *A. afarensis* has consistently been recognized as the ancestor of *A. africanus,* but views vary about whether its connections with *Homo* are direct (path 1b in the figure) or indirect—*i.e.,* via *A. africanus* (path 1a). *A. africanus* was once regarded as the ancestral hominid, but few contemporary researchers would subscribe to this hypothesis. Studies which have taken characters and compared their expression, one by one, with those in other



1a, 1b Alternative ancestry of *Homo habilis*
2a, 2b Alternative ancestry of *Australopithecus robustus*
3a, 3b Alternative ancestry of *Australopithecus boisei*
4a, 4b, 4c Alternative ancestry of "robust" australopithecine clade

Figure 14: Tentative phylogenetic scheme for early hominid evolution. A solid line between species indicates general agreement on a single evolutionary pathway; dotted lines signify that more than one pathway has been proposed.

taxa have demonstrated that *A. africanus* has affinities with both *Homo* and the robust australopithecines. This mixture of traits has suggested to some workers that *A. africanus* is the common ancestor of both *Homo* and *A. robustus* (paths 1a and 2a), but others see a preponderance of either *Homo* (path 1a) or robust australopithecine (path 2a or paths 4a, 2b, and 3b) traits. The discovery in the mid-1980s of the KNM-WT 17000 cranium at the West Turkana site has further sharpened the debate. If, as is generally believed, the robust australopithecines (*A. robustus* and *A. boisei*) represent a separate branch, or clade, of the evolutionary tree, then the evidence gathered from the study of KNM-WT 17000 reduces the probability that *A. africanus* can be included in that grouping. Conversely, if *A. africanus* is judged to show robust australopithecine specializations, then the hypothesis must seriously be entertained that the East and southern African forms of robust australopithecine are similar because of convergence and not because of shared ancestry (paths 2a and 3b).

The KNM-WT 17000 cranium

The surest way to adjudicate on the claims of these rival hypotheses lies in the exploration of the detailed morphology that underlies the various characters alleged to link australopithecine taxa. Whereas characters inherited from a common ancestor are likely to be similar in most details, it is predictable that phylogenetically misleading, or convergent, characters will not show such detailed similarities. (B.Wo.)

Homo habilis

The extinct species of the genus *Homo* called *Homo habilis* inhabited parts of sub-Saharan Africa at least two million years ago. In 1959 and 1960 the first *H. habilis* fossils were discovered at Olduvai Gorge in northern Tanzania. These consisted of several teeth and a lower jaw associated with fragments of a cranium and some hand bones. As more specimens were unearthed workers began to realize that the hominids they represented were anatomically different from *Australopithecus.* Formal announcement of this and other discoveries was made by L.S.B. Leakey, Phillip Tobias, and John Napier in 1964. They described increased cranial capacity and comparatively smaller premolar and molar teeth as factors leading to the designation of the fossils as *H. habilis* and also suggested that the hand was capable of fine manipulation; *Homo habilis* thus seemed to foreshadow conditions seen in *Homo erectus* and in later humans.

FOSSIL EVIDENCE

Apart from the jaw, cranial bones, and hand thought to represent a juvenile individual (Olduvai Hominid, or OH, 7) and recovered from Bed I deposits, several additional fossils from Olduvai have been ascribed to *Homo habilis.* At Bed II, at a site somewhat higher in the deposits than Bed I, pieces of another thin-walled cranium along with upper and lower jaws and teeth came to light in 1963. Just a month later a third skull was found in Bed II, but these bones had been trampled by cattle after being washed into a gully. Some of the teeth survived, but the cranium was broken into many small fragments; only the top of the braincase has been pieced back together. The two skulls from Bed II are numbered OH 13 and OH 16, and both are mentioned in the report prepared by Leakey, Tobias, and Napier in 1964.

The Olduvai and Koobi Fora finds

G. Philip Rightmire

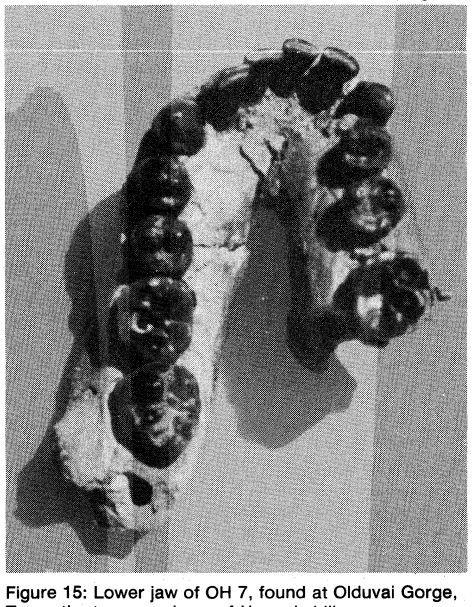


Figure 15: Lower jaw of OH 7, found at Olduvai Gorge, Tanz.; the type specimen of *Homo habilis.*

Since 1964 more material has been discovered, not only at Olduvai but at other African localities as well. One intriguing specimen is OH 24. This cranium is more complete than others from Olduvai ascribed to *H. habilis.* Because some of the bones are crushed and distorted, however, the face and braincase are warped and provide less anatomical information than they otherwise would. OH 24 is said to differ from *Australopithecus* in brain size and dental characteristics, but there are resemblances to the australopithecines of southern Africa in other features, such as face form. Partly because the fossil is damaged, complete agreement concerning its significance has not been reached.

Important discoveries made in the Koobi Fora region of northern Kenya include the famous cranium numbered KNM-ER 1470. As in the case of OH 16, this specimen had been broken into many fragments, which could be collected only after extensive sieving of the deposits. Some of the pieces were then fitted into the reconstruction of a face and much of a large cranial vault. Brain volume can be measured rather accurately and is about 750 cubic centimetres (cc; 46 cubic inches). This evidence prompted Richard Leakey to describe KNM-ER 1470 as one of the oldest undoubted representatives of the genus *Homo* to be unearthed in East Africa. Some other features of the braincase are *Homo*-like, and many workers have accepted this opinion. At the same time, it is apparent that the facial skeleton is relatively large and flattened in its lower parts. In this respect, the Koobi Fora specimen resembles *Australopithecus* anatomically.

Among other key finds from the Koobi Fora region are KNM-ER 1813 and ER 1805. The former, which is most of a cranium, is smaller than ER 1470 and resembles OH 13 from Olduvai in many details, including tooth size and morphology. The latter skull exhibits some peculiar features. Although the braincase of ER 1805 is close to 600 cc in volume and is thus expanded moderately beyond the size expected in *Australopithecus,* a bony crest is formed along the top of the vault. This sagittal crest is coupled with another, more massive crest, oriented transversely across the rear of the skull. These ridges indicate that the temporal muscles (which function in chewing) and also the neck muscles were powerfully developed. A similar if more exaggerated pattern of cresting appears in what are referred to as robust australopithecines but not in *Homo.* Other features of ER 1805, however, are *Homo*-like. As a result, there has been disagreement among anatomists regarding the hominid species to which this individual should be assigned. Despite its anomalies, ER 1805 is probably best discussed along with other specimens grouped as *Homo habilis.*

Several mandibles resembling that of OH 7 have been recovered from the Koobi Fora area, and teeth that may belong to *H. habilis* have been found further to the north, in the Omo River valley of Ethiopia. Some additional materials, including a badly broken cranium, are known from the cave at Swartkrans in South Africa. At Swartkrans the fossils are mixed in the Member 1 deposits with many other bones of robust australopithecines. An early species of *Homo* may also be present in the Member 5 breccias at Sterkfontein, not far from Swartkrans. Here again the remains are fragmentary and not particularly informative.

Finds at other sites

A more recent and valuable discovery has been reported from Olduvai Gorge: in 1986 a jaw with teeth, cranial parts, and pieces of a right arm and both legs were located low in Bed I. The bones seem to represent one individual,

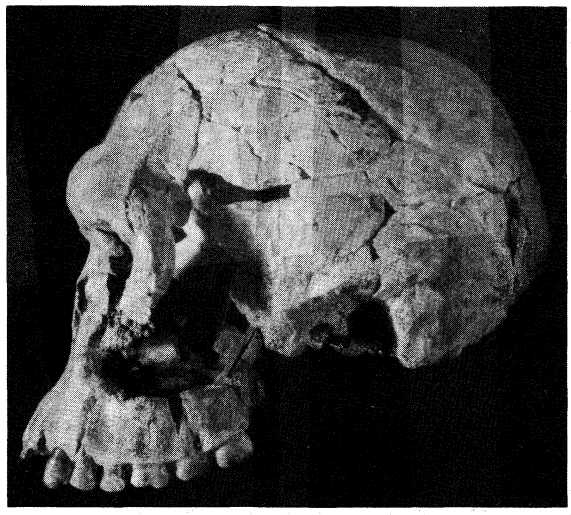


Figure 16: Left side view of the KNM-ER 1813 cranium, found at Koobi Fora, Kenya.
G. Philip Rightmire

called OH 62. Although the skull is shattered, enough of the face is preserved to suggest similarities to early *Homo*. The find is especially important because of the associated limbs, which show that OH 62 is a very small hominid indeed. The arm is long relative to leg length so that this individual has body proportions which differ dramatically from those of more modern hominids.

DATING THE FOSSILS

Fossils from Olduvai that are attributed to *Homo habilis* come from Beds I and II. Several approaches have been used to obtain dates for these stratigraphic levels. One of the most tested and most reliable techniques, potassium–argon dating, measures the amount of argon relative to radioactive potassium in samples of volcanic rock or minerals. Volcanic materials suitable for dating often occur as tuffs deposited along with other sediments in the Olduvai sequence. Other methods, such as magnetic stratigraphy and measurement of sediment thicknesses, also yield clues to the age of deposits. These approaches are subject to uncertainty, but a reasonably secure time scale for Olduvai has been developed. The oldest remains, including OH 24, were found just below a tuff that is about 1,850,000 years old. Other individuals from Bed I such as OH 7 and OH 62 are not quite so ancient. The youngest Olduvai skull that is representative of early *Homo* is OH 13, from Middle Bed II. No radiometric date for this level is available, but OH 13 is probably about 1.5 million years in age.

In the Koobi Fora region a number of the important fossils have been located near a level of volcanic ash that also contains stone tools. This ash bed, known as the KBS tuff, was sampled and dated initially by potassium–argon to about 2.6 million years. When ER 1470 was found several metres below this tuff in 1972, it was thought that the new cranium must document *Homo* in the record at a time well before the Olduvai deposits had accumulated. This assumption was soon questioned on the basis of other evidence, however, and before 1980 it was clear that the age of the KBS tuff had been overestimated. A series of potassium–argon determinations done subsequently has yielded a date of 1,880,000 years. ER 1470 and other *H. habilis* specimens recovered below this ash layer, therefore, must be close to two million years old. Remains collected above the tuff are somewhat younger, but probably none of the Koobi Fora fossils is as recent as OH 13. Dating evidence from East Africa thus suggests that *H. habilis* lived for half a million years or so before giving way to later *Homo* species.

BODILY STRUCTURE OF HOMO HABILIS

Olduvai and Koobi Fora fossils have allowed researchers to make some determinations about the anatomy of early *Homo*. It is clear that the braincase of *H. habilis* is larger than that of *Australopithecus*. The original finds from Olduvai Gorge include two parietal bones from OH 7. An incomplete brain cast that was molded when the parietals were put together to form a partial cranium has been used to estimate total brain volume, and the result, after correction for the juvenile status of OH 7, is about 680 cc. Less-direct methods can also be employed to assess the endocranial volume of OH 7, and most give figures close to 700 cc. A brain cast from ER 1470, which has a more complete cranium, can be measured directly from water displacement; its volume, as mentioned above, is about 750 cc. One or two additional skulls from the East Turkana (Koobi Fora) region that are fragmentary appear to be about the same size as ER 1470. Other individuals—such as ER 1813, which has a cranial capacity of only about 510 cc—are much smaller. Thus, brain sizes ranging from slightly more than 500 cc to nearly 800 cc seem to characterize *H. habilis*.

Increase in brain size

The craniums by and large have thin walls and a rounded, rather than low and flattened, vault; they do not have the heavy crests and projecting browridges characteristic of later *H. erectus*. As with *Australopithecus* and archaic *Homo* the occipital of these specimens is flexed, but unlike these others the transverse torus does not protrude. The area of the lower occiput covered by the nuchal (neck) muscles is much smaller than that of *H. erectus*. The underside of the cranium is shortened from the back of the palate to the occipital bone, as in all later *Homo* species. This is an important contrast to the condition exhibited by the gracile australopithecines, where the cranial base is relatively narrow and elongated.

The facial bones of several specimens are at least partly preserved, and facial proportions vary considerably. One of the Olduvai hominids, OH 24, seems similar anatomically to *Australopithecus*, having prominent cheekbones and a flattened nasal region. This gives the central region of the face a depressed, or "dished," appearance, and the upper part of the nasal profile is obscured by the cheek when the specimen is viewed from the side. Such hollowing of the face is characteristic of some South African australopithecines but is not seen in later *Homo*. The facial skeleton of ER 1470 is large relative to the braincase and shows flattening below the nose, again *Australopithecus*-like features. The walls of the nasal opening, however, are slightly everted, and there is at least an indication that the nose stands out in more relief than would be expected in australopithecines. The face of ER 1813 is still more modern in form: the edge of the nasal opening is thin and flared outward, and this region is prominent relative to the cheeks on either side.

The anterior teeth of *H. habilis* are not much different in size from those of *Australopithecus*, but the premolar and molar crowns—particularly in the mandible—are narrower. The jaw itself may be quite heavily constructed and overlap in size with mandibles of gracile australopithecines. This is the case for OH 7 from Olduvai and also for at least one specimen from Koobi Fora. Other jaws are smaller but still robust, in the sense of being thick relative to height. For example, the mandible of OH 13 is similar in many respects to that of *H. erectus*, and this individual might have been called *H. erectus* if its jaw had not been found with the small, thin vault bones that mark it as different from that species.

Only a few postcranial parts have been discovered. Some limb bones from Olduvai and Koobi Fora have been grouped tentatively with *H. habilis* on the strength of general anatomic similarity to later humans. These fossils, however, are not associated with teeth or skulls, and it is probably not appropriate to use them as the basis for describing early *Homo*. One individual for which body parts are more fully represented is OH 62 from Olduvai. Arm and leg bones of OH 62 are fragmentary and must be studied in greater detail, but a preliminary report by the discoverers suggests that the arm is relatively long. The skeleton may be similar in its proportions to small species of *Australopithecus*. Probably OH 62 walked bipedally as efficiently as other early hominids, but this diminutive individual was unlike later humans in many respects.

Another important specimen is the immature hand of OH 7. These bones found with the parietals are still ape-

like in some features, but it is almost certain that the individual from which they came could manipulate objects with precision. Stone artifacts and early *Homo* fossils have been found in the same levels at Olduvai and other sites. These tools are called the Oldowan industry, and though they are crude they do indicate that *H. habilis* could shape stone.

Use of tools

BEHAVIORAL INFERENCES

The stone tools and unused waste materials (mainly crude chopping tools and sharp flakes) left by *H. habilis* provide important clues about the behaviour of these early humans. Olduvai Gorge has been a rich source of Oldowan tools, where they are found in several levels of Bed I, often in association with animal fossils. Originally, the occurrence of artifacts with bones was interpreted to mean that *H. habilis* hunted animals, brought them to their living sites, and butchered the carcasses with the Oldowan implements, but through other studies it is now known that the situation is more complicated than this. Assemblages such as those found at Olduvai can be created through various means, not all of which are related to hominid activities.

Further study of the Olduvai material has indicated that *H. habilis* did use animal products. With the aid of a scanning electron microscope it has been shown that cut marks on some of the Bed I bones must have been made by stone tools, but this does not prove that animals were hunted. Analysis of Olduvai animal fossils shows that marks were made by either rodent or carnivore teeth and by cutting, the indication being that at least some of the animals were killed by nonhominid predators. In all likelihood, the hominids at Olduvai could obtain larger carcasses only after the animals had been killed and partially eaten by other predators. *H. habilis* may have hunted small prey, such as antelopes, but they were also scavengers.

It is debatable whether or not the Olduvai sites were home bases. Nothing recovered indicates that people lived where the animal bones accumulated, and presumably such areas were dangerous since they undoubtedly attracted numerous predators. These sites may have been caches of stone tools and raw materials that were established in areas convenient for the rapid processing of animal parts. Thus, where the hominids lived or whether their social structure was prototypical of later hunter-gatherers remains unknown, although *H. habilis* must have engaged in cultural activities.

Language ability

Whether or not early *Homo* had acquired language is another fundamental question, and the indirect evidence on this issue has been variously interpreted. It is the belief of some anatomists that endocranial casts of *H. habilis* fossils indicate that the regions associated in modern humans with speech are enlarged. Other workers have disagreed with this assessment, particularly since the number of braincases preserved well enough to make detailed casts is small. Anthropologists have based their interpretations on the archaeological record. According to some, the crude Oldowan artifacts indicate the ability to use language. Critics of this view assert that the Oldowan industry represents only opportunistic stonework and that the later Acheulean tools of *H. erectus*—because they are more carefully formed and are often highly symmetrical—indicate this later hominid species to have been the first to use symbols and language. One of the problems with this theory is that no clear link between technological and linguistic behaviour has been established; even the more sophisticated tools could have been made by nonspeaking hominids. Thus, it is not certain when *Homo* developed the linguistic skills that characterize modern humans.

EVOLUTIONARY IMPLICATIONS

The general interpretation of the fossil evidence is that *Homo habilis* is anatomically different from *Australopithecus* and that it represents the beginning of the trends which characterize human evolutionary history. One of the hallmarks is the expansion of the brain. Some of the Olduvai and Koobi Fora hominids clearly have a larger cranial capacity than that of *Australopithecus* specimens, and the capacity increases progressively with *H. erectus*, archaic *H. sapiens*, and modern humans. *H. habilis* is also thought to exhibit the origins of such other trends as smaller dentition and changes in facial structure, especially of the nasal region. In addition, the argument has been made that *H. habilis* could fashion simple tools and could communicate verbally. It is fairly certain that these early hominids had the technological skills to make the Oldowan artifacts, but it is not clear that this stoneworking ability alone meant that they could speak; there is simply too little evidence to make such an inference.

The theory that *H. habilis* is intermediate between relatively primitive Pliocene *Australopithecus* and more advanced *Homo* of the Pleistocene appears to be generally accurate, but several aspects of this view can be challenged. Although *Homo habilis* is still a rather poorly known species, it is becoming clear that there are anatomic differences within the East African assemblages. Some of the newer discoveries, such as ER 1470 from Koobi Fora, have confirmed the expectation that early *Homo* crania should be relatively large, with rounded occiputs and shortened bases. Other fossils have proved less easy to assign to *H. habilis*, and there has been considerable controversy over OH 24 from Olduvai and ER 1805 and ER 1813 from Koobi Fora. These braincases are considerably smaller, and it is frequently suggested that this variation may be due to sex: OH 7 and ER 1470 are said to be male, while OH 24 and ER 1813, with their smaller craniums, female. But there are also differences in shape as well as size, and several of the smaller skulls depart from the morphology of large-brained *H. habilis* in ways that are not obviously related to sex. The facial affinities between OH 24 and *Australopithecus* have been mentioned above, and ER 1805 exhibits cranial cresting patterns that are unlike those seen in other *Homo*. There is the possibility that two taxa, rather than one sexually dimorphic group, are actually represented by the fossils.

Anatomic differences

Placing the specimens in two groups, however, means that both must be fitted into a scheme of hominid phylogeny. One interpretation assigns specimens with smaller craniums (including OH 24) to a "gracile" species of *Australopithecus*. According to this scenario, only the larger skulls (including OH 7 and ER 1470) represent early *Homo* evolution. Others believe that early human populations were more diverse than had been recognized, questioning the notion that all species of *Homo* form a simple linear progression; although these workers recognize two separate taxa, they prefer to lump both in the genus *Homo*. In this view, two species may have lived contemporaneously two to 1.5 million years ago, but only one was the direct ancestor of *H. erectus*. Perhaps it was the large-brained form that evolved further, while the smaller hominid became extinct. What is certain is that getting a clearer understanding of the history of these first humans requires further paleoanthropological research. (G.P.Ri.)

Homo erectus

Homo erectus, the first generally recognized human species, most likely originated in Africa, and it quite possibly evolved from *Homo habilis*. *H. erectus* seems to have been restricted to the African tropics for several hundred thousand years, but eventually these people gradually migrated into Asia and probably into parts of Europe. This history can be documented directly from the many sites that have yielded fossil remains of *H. erectus*. Other localities from which animal bones and stone tools have been recovered indicate that this species was present, although there is no evidence of the people themselves. *H. erectus* seems to have flourished until sometime in the Middle Pleistocene—perhaps 300,000 years ago—before giving way to early representatives of *Homo sapiens*.

FOSSIL EVIDENCE

The first fossils that came to be attributed to the species were discovered by a Dutch army surgeon, Eugène Dubois, who began his search for ancient human bones in Java (now part of Indonesia) in 1890. Dubois found his first specimen in that year, and in 1891 a fine skullcap was unearthed at Trinil on the Solo River. Considering its prominent browridges, retreating forehead, and flexed oc-

The Java finds

ciput, Dubois concluded that the Trinil cranium showed anatomic features intermediate between those of humans—as they were then understood—and those of apes. Several years later, near where the skull was discovered, he found a remarkably complete and modern-looking femur (thighbone). Since this long, straight bone was so much like a modern human femur, Dubois decided that its owner must have walked erect. He adopted the name *Pithecanthropus,* which had been coined earlier by the German zoologist Ernst Haeckel, calling his discoveries *Pithecanthropus erectus,* or "Upright Ape-Man." Only a few other limb fragments turned up in the Trinil excavations, and it would be some three decades before more substantial evidence of these archaic people appeared. Most paleontologists now regard all of this material as *Homo erectus,* and the name *Pithecanthropus* has been dropped.

Asian fossils. Subsequent discoveries gradually established the case for a new and separate species of fossil hominid. At first, these discoveries were centred largely in Asia. At several different places in Java, essentially similar fossils were found: the sites other than Trinil are Kedung Brubus, Modjokerto (Mojokerto), Sangiran, Sambungmatjan (Sambungmachan), and Ngandong. Another series of finds was made in China, especially in the caves of Chouk'ou-tien (Zhoukoudian) near Peking, although virtually all of the remains from there subsequently were lost during the Sino-Japanese War late in 1941. Newer discoveries have since been made in the caves, while three new Chinese sites—at Kung-wang-ling (Gongwangling) and Ch'en-chia-wo (Chenjiawo) in the Lan-t'ien (Lantian) district of Shensi province, and at Ho-hsien (Hexian) in Anhwei province—have yielded remains attributable to *H. erectus.*

By the end of World War II the pattern of early discovery had given rise to an idea that *H. erectus* was a peculiarly Asian expression of early humans. Subsequent discoveries in Africa served to change this view, and it came to be realized that Europe, too, may have harboured *H. erectus.*

African fossils. In Africa in 1954–55, excavations at Ternifine, east of Mascara, Alg., yielded remains whose nearest affinities seemed to be with the Chinese form of *H. erectus.* Other hominid fragments from northwestern Africa—parts of a skull found in 1933 near Rabat in Morocco and jaws and teeth from Sīdī ʿAbd ar-Raḥmān (Sidi Abderrahman; 1954) in Morocco—show features reminiscent of *H. erectus,* though they are rather more advanced in structure than those of Ternifine and of Asia. Another fossil probably related to *H. erectus* is a cranium found in 1971 at Salé, Mor. Although nearly all of the face and part of the frontal bone have been broken away, it is an important specimen.

Finds at Olduvai Gorge

Some of the more convincing evidence for the existence of *H. erectus* in Africa came with the discovery in 1960 of a partial braincase at Olduvai Gorge in Tanzania. This fossil, called OH 9, was picked up by L.S.B. Leakey in deposits of Bed II, the second of four numbered beds (or layers) identified in the Olduvai stratigraphic sequence, beginning with Bed I (the oldest) and ending with Bed IV (consisting of more recent sediments). Additional cranial remains, jaws, and limb bones of *H. erectus* were later discovered in Beds III and IV and the overlying Masek levels.

Much of the Olduvai material is fragmentary, but gaps in the knowledge of *H. erectus* in East Africa have been filled to some extent through finds made by Richard Leakey. Since 1970 a number of valuable fossils have been unearthed at localities on the eastern shore of Lake Turkana (Lake Rudolf) in northwestern Kenya, now com-

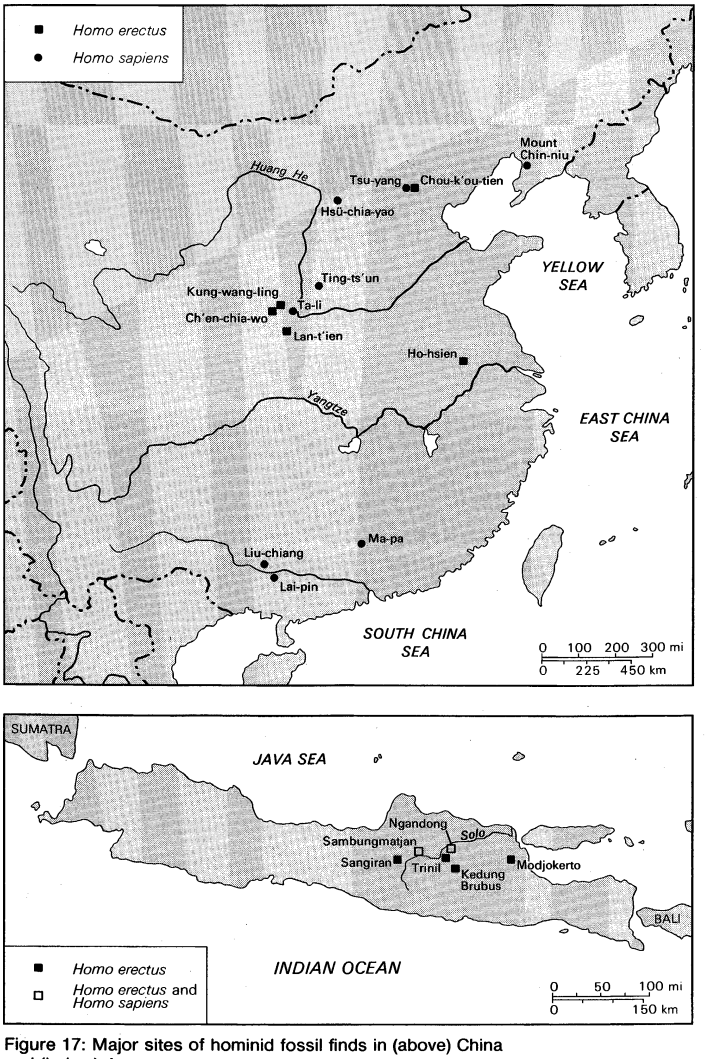


Figure 17: Major sites of hominid fossil finds in (above) China and (below) Java.

monly referred to as the Koobi Fora sites. Included in these assemblages are the remains of *Australopithecus* and probably some representatives of the species *H. habilis.* Of several specimens that clearly do belong to *H. erectus,* one cranium from Koobi Fora (KNM-ER 3733) is quite complete and well preserved; it is likely to be one of the most ancient *H. erectus* fossils discovered anywhere in Africa. Other significant finds at Koobi Fora include a nearly intact skeleton (ER 1808), although it has been shown to have come from a diseased individual. Another nearly complete skeleton (designated KNM-WT 15000) found at Nariokotome (West Turkana), a site on the northwestern shore of Lake Turkana, is of an adolescent male.

European fossils. The realization that Africa as well as Asia was apparently peopled by a form of mankind classifiable as *H. erectus* led to a reexamination of some of the earliest hominid fossils from Europe. An isolated mandible (lower jawbone) had been found in a sandpit just north of Mauer, close to Heidelberg, Ger., in 1907. Although it had been given a variety of names over the years, its exact affinities to other fossils remained uncertain, because no associated cranium was found. Since then, a number of investigators have come to regard the Mauer mandible as representing a member of the species *H. erectus.* Although its geologic age is perhaps comparable with that of the Chou-k'ou-tien hominids in China, this skeletal fragment from Europe shows more modern structural features than do the Asian and African jaws of *H. erectus.* The exact significance of these features in the Mauer jaw is still being debated; they could be the mark of an individual variant, highlighting the fact that it is not yet known how variable in bony structure *H. erectus* was. Alternatively, the Mauer specimen could represent a subspecies or race of *H. erectus* that is slightly more advanced in anatomic structure than are the African and Asian populations; or it is possible that the Mauer individual could have been a member of a very early population of *H. sapiens.* Evidence for the latter view has been provided by fossils discovered in Hungary.

The Mauer mandible

In 1965, remains of two individuals—a child and an adult—were found in a travertine quarry at Vértesszőllős, Hung., about 30 miles (50 kilometres) west of Budapest. The remains of the child are milk teeth, and enough of them are preserved to show affinities with the Chinese *Homo erectus* of Chou-k'ou-tien. The adult is represented by a large part of an occipital bone (at the back of the head) from a large-brained skull. While showing some features reminiscent of *H. erectus,* the general form and the capacity of the Hungarian cranium also suggest an affinity with an early branch of *H. sapiens.* In fact, it appears to be related to somewhat later Middle Pleistocene skulls (perhaps 200,000 years old) from Europe, such as those of Swanscombe and Steinheim, which are accepted as early members of *H. sapiens.* Since such twofold affinities are exhibited by the Vértesszőllős group of remains, authorities differ as to whether to call the population they represent *H. erectus* or *H. sapiens.* The same uncertainty applies to some of the fossils found in North Africa; the Sīdī ʿAbd ar-Raḥmān and Rabat remains are regarded by some experts as late surviving members of *H. erectus* and by others as forms transitional between *H. erectus* and *H. sapiens.*

The European fossils that are of equivalent age to those of Chou-k'ou-tien and the later northwestern African localities—including the jaw from Mauer; the teeth and occipital bone from Vértesszőllős; and some skull parts, teeth, and postcranial remains discovered more recently at Arago cave (near the village of Tautavel in southwestern France)—appear to have more structural features in common with early *Homo sapiens* than do their Asian and African contemporaries. This is also true of several other European finds that are somewhat younger (that is, of later Middle Pleistocene antiquity). Pieces of braincase and an upper molar tooth from travertine deposits at Bilzingsleben, E.Ger., show some resemblances to *H. erectus,* but the skull fragments are more comparable to those of a remarkably complete hominid recovered in 1960 from a limestone cave near Petralona, Greece. Detailed study of the anatomy of this latter hominid suggests that it shares with *H. sapiens* some traits that are not characteristic of

The Petralona skeleton

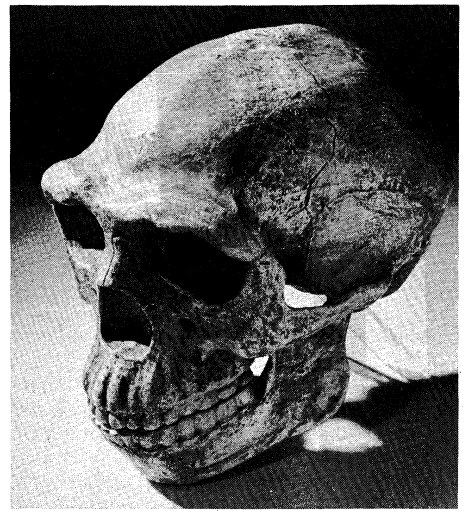


Figure 18: Reconstructed skull of Peking man (*Homo erectus pekinensis*).
By courtesy of the University Museum of Archaeology and Ethnology, Cambridge

H. erectus. There is still no firm consensus regarding the classification of these later Middle Pleistocene populations of Europe. It may be safest to conclude that the case has not yet been convincingly established for the existence in Europe of *H. erectus* as known from Asia and Africa.

DATING THE FOSSILS

To reconstruct the position of *Homo erectus* in hominid evolution, it is essential to define his place in time as precisely as possible. Modern developments in such disciplines as physics have placed at the disposal of the paleoanthropologist a variety of techniques that permit increasingly accurate assessments of the absolute age of fossils. Many of these methods are based upon the effectively constant (or absolute) rate at which radioactive isotopes of such elements as potassium and argon decay. When the newer methods cannot be applied, investigators may still ascribe a relative age to a fossil. This can be done by noting the contents of the layer of rock or the deposit in which the fossil was found; a layer containing, for instance, evidence of the remains primarily of extinct animals is probably older than one containing signs of predominantly recent or still living forms.

Such lines of evidence have led to the tentative conclusion that the species *Homo erectus* flourished over a long interval of Pleistocene time. The fossils recovered at the Koobi Fora sites may be about 1.6 million years old if radiometric dates obtained from associated volcanic materials are correct. Some of the remains from Olduvai Gorge are also ancient, and OH 9 is probably about 1.2 million years old. Many of the other older-appearing fossil hominids unfortunately lack absolute dates; but, by relative dating from several approaches, it would seem probable that the specimens from Sangiran and Modjokerto in Java (more especially the fossils from deposits of the Putjangan [Pucangan] beds) and perhaps from one of the Lan-t'ien localities in China are among the earlier representatives of *H. erectus.* On the other hand, the youngest hominids accepted unequivocally as *H. erectus* would seem to be the group from Ternifine in Algeria, Chou-k'ou-tien in China, and Trinil in Java.

Until more complete evidence is available it is reasonable to suggest 1.6 million to perhaps 250,000 years ago as a time range for *Homo erectus.* For the most part, fossils dated earlier than this are the remains of *H. habilis.* This species is best known from Olduvai Gorge and Koobi Fora in Africa, and the oldest specimens from there seem to be about 1.8 to two million years old. On the other hand, there is a group of later specimens that show some features of *H. erectus* but are commonly regarded as transitional forms or as members of *H. sapiens;* these include later Middle Pleistocene specimens from Europe (discovered at sites such as Bilzingsleben, Petralona, and

Transitional forms

Montmaurin), from northwestern Africa (Salé, Sīdī 'Abd ar-Raḥmān, and Rabat), and from Asia (the Ta-li find of 1978). Other later forms suggest that *H. erectus* had given rise to several regionally distinct forms, or subspecies, of archaic *H. sapiens,* represented by late Middle Pleistocene or early Late Pleistocene fossils from Africa (Kabwe/Broken Hill, Elandsfontein [Hopefield/Saldanha], Cave of Hearths, Lake Ndutu, Omo, Bodo) and Europe (Swanscombe, Steinheim, Biache, Ehringsdorf, La Chaise). Thus, the problem of recognizing populations as belonging to *H. erectus* becomes more difficult; the boundaries of the species become blurred. These are the transitional zones in which a predecessor species seems to have been grading imperceptibly into its evolutionary product, *H. erectus,* and in which *H. erectus* apparently was undergoing further evolutionary change into its descendant species, *H. sapiens,* to which modern humans belong.

BODILY STRUCTURE OF HOMO ERECTUS

Much of the fossil material discovered in Java and China consists of cranial bones, mandibles, and teeth. The few broken limb bones found at Chou-k'ou-tien have provided little information. It is possible that the complete femur excavated by Dubois at Trinil is more recent in age than the other fossils found there and not attributable to *Homo erectus.* It comes as no surprise, therefore, that the greatest descriptive emphasis has been on the shape of the skull and not on the postcranial finds. The continuing discoveries in Africa—particularly at the Olduvai and Lake Turkana sites—have yielded a more complete picture of *H. erectus* anatomy.

The cranium of *H. erectus*—with its low profile and average endocranial (interior) capacity of less than 1,000 cc (60 cubic inches)—is distinctly different from that of other humans. The average endocranial capacity of modern *H. sapiens,* for example, is 1,450 cc, although the range for recent humans is appreciable, perhaps 1,000 to 2,000 cc. The upper part of the range for *H. erectus,* extending to more than 1,200 cc, overlaps with the lower values expected for *Homo sapiens.*

Table 2: Average Capacity of the Braincase in Fossil Hominids

hominid	number of fossil examples	average capacity of the braincase (cc)
"Gracile" *Australopithecus*	6	440
"Robust" *Australopithecus*	4	519
Homo habilis	4	640
Javanese *Homo erectus*	7	883
Chinese (Peking) *Homo erectus*	5	1,043
Homo sapiens	7	1,450

Some difference in estimated brain size is apparent between the Javanese and the Chou-k'ou-tien populations of *Homo erectus.* Thus, for seven Javanese craniums, the average is 883 cc, with a range from 750 to 1,030 cc; while for five Chou-k'ou-tien craniums, the capacity ranges from 915 to 1,225 cc and averages about 1,043 cc. That is, the mean capacity in the Peking fossils of *H. erectus* exceeds that of the Javanese by about 160 cc. Investigators note with interest that the cranium from Kung-wangling (also Chinese), which comes from an earlier period than do those of Chou-k'ou-tien and is an approximate contemporary of the earlier fossils from Java, shares with the Javanese group a smaller cranial capacity (780 cc).

Differences in brain size

Theoretically, the difference in brain size between the two groups of Asian fossils of *H. erectus* may be the consequence of further evolutionary increase in brain size in later populations of *H. erectus.* Alternatively, it may simply be interpreted as representing differences in average features between two different "races" or subspecies of *H. erectus.* Several African values are available, and in the case of the Koobi Fora and Olduvai individuals, these range from about 850 cc (ER 3733) to 1,067 cc (OH 9).

While the cranial capacity of *H. erectus* falls short of that of *H. sapiens* by about 570 cc on the average for the Javanese group and about 400 cc for the Peking group, *H. erectus,* in turn, exceeds the australopithecine capacities by an average of about 440 cc in the case of the Javanese fossils and about 600 cc for the Peking group. Thus, the cerebral gap between *Australopithecus* and *H. erectus* is slightly greater than that between *H. erectus* and *H. sapiens.* Into the former gap fit the cranial capacities of the group of fossils known as *Homo habilis.* Clearly, the last word has not been written on their affinities. The possible evolutionary place of the cranial capacity of *H. erectus* can be seen in Table 2 (above), but it perhaps emerges more clearly in Figure 19.

From W. Howells, *Evolution of Genus Homo* (Benjamin/Cummings Pub. Co., copyright © 1973), p. 47; reproduced by permission of Benjamin/Cummings Pub. Co.

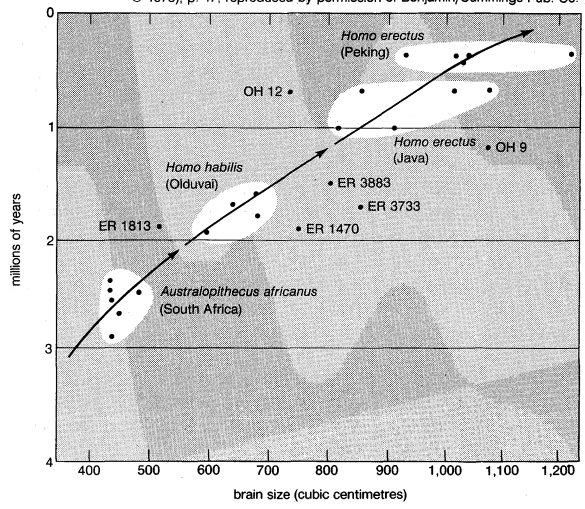


Figure 19: The increase in hominid cranial capacity over time.

Apart from their characteristically small capacity, the skulls of *H. erectus* show a number of distinctive features. The face, which is preserved in only a few specimens, is massively constructed and projecting in its lower parts. The bone forming the wall of the nose is thinner and more everted than in earlier *Homo* or *Australopithecus,* and the nasal profile is not too different from that seen in modern humans. The braincase is low, with sides that taper upward, and the bones of the cranial vault are thick. Over the eye sockets is a strongly jutting browridge, called a supraorbital torus. There is a flattened forehead, and the front part of the cranium immediately behind the supraorbital torus is appreciably constricted from side to side. A low ridge or keel of bone may extend from the frontal bone onto the parietals (side walls) along the midline, and there tend to be strongly developed crests in the mastoid region. The broad-based skull has a sharply curved occiput and a markedly thickened shelf of bone (occipital torus) that divides the upper and lower occipital surfaces. The area of neck-muscle attachment below the occipital torus is much larger than in *H. habilis* or *H. sapiens.* Other distinguishing features in *H. erectus* can be found on the underside of the cranium, especially at the joint of the mandible. The mandible itself is deep and robust and lacks chin development. The teeth are on the whole larger than those of *H. sapiens.*

Postcranial fossils

The femur is the most commonly recovered postcranial fossil. Apart from the Trinil specimen, the affinities of which are open to doubt, a number of femurs have been found at Chou-k'ou-tien, and more have been recovered from sites in Africa. Several notable characteristics have been observed in the *H. erectus* femur. The form of these bones resembles that of modern humans, and *H. erectus* must have walked upright efficiently. On the other hand, the construction of the bones is robust, a condition also seen in other skeletal members. This robusticity suggests that the life-style of *Homo erectus* was physically demanding.

Limb bones also supply information about the size of *H. erectus.* Size is important to any hominoid species, since it influences behaviour and various aspects of anatomy, including bodily proportions. One measure of size is stature,

or height. The femurs found at Chou-k'ou-tien and Koobi Fora are too broken to yield a good estimate of the height of these individuals, but accurate measurements of the skeleton numbered KNM-WT 15000 from Nariokotome have been made. Although he was not fully grown, it is thought that the Nariokotome boy would have reached about 160 centimetres in height, which is close to the stature expected for modern adult males.

The total pattern of the bodily structure of *H. erectus,* as preserved in the bones, is rather different from that of *H. sapiens.* Parts of the postcranial skeleton are robust but otherwise generally comparable to those of modern humans. The brain is relatively small, though not so small as that of *Australopithecus* and *H. habilis.* In addition, in this hominid's thick skull bones and extraordinarily developed eyebrow ridges and occipital torus, some investigators say they see unique, specialized features, not characteristic either of its presumed ancestors or of apes and not pointing to *H. sapiens* as the direction of subsequent evolution.

Some scientists even infer that these last traits show *Homo erectus* to have specialized so far off the modern human line that it could not have been ancestral to *Homo sapiens.* It is notable that the more ancient *Australopithecus* had thin skull bones and only modest protuberances on the cranium; while the later *H. sapiens* also tends to have thin skull bones with a marked diminution in the size of crests and ridges sculpted on the surface of the cranium. It is often said that only two choices are open in interpreting this situation: either *H. erectus,* with a thick cranium and large adornments on the skull, could not have been on any direct evolutionary line from *Australopithecus* to *H. sapiens;* or else *H. erectus* evolved specialized features from *Australopithecus* and then lost them again (or underwent a type of evolutionary reversal) to produce the thin, smooth cranium of modern *H. sapiens.*

Yet, the two choices offered for solving the problem are judged by some authorities to be oversimplified. It has been observed that the exponents of either view probably fail to take into account that there is very little evidence about the variability of these features—cranial thickness and external embellishments of the skull—among members of even one population of *H. erectus,* let alone among different populations of *H. erectus* dispersed through two or three large continents. Then, too, practically nothing is known about the climatic or ecological conditions under which cranial thickening occurred, or of the effect on skull growth of the brain enlargement that was so striking a feature of the evolutionary advance of earliest Pleistocene humans to the people of the later Middle Pleistocene and of the Late Pleistocene. These and many other questions need answers before *H. erectus* can be written off as an ancestor of *H. sapiens.* In the meantime, another hypothesis that meets most of the available evidence is that *H. erectus* was in the process of evolving from pre-*Homo erectus*—probably *Australopithecus* and *Homo habilis*—to post-*Homo erectus;* that is, to *Homo sapiens.* In most details, the bodily structure of *H. erectus* fulfills what might have been predicted for an intermediate between *Australopithecus* and *H. sapiens.*

Unanswered questions

BEHAVIORAL INFERENCES

To understand this species of fossil humans more completely requires looking past their bones and beyond their dispersal in time and space. What evidence is there that bears upon the behavioral or cultural pattern of their daily existence?

First, the discovery sites themselves may throw light on this question. At Chou-k'ou-tien, the remains of *H. erectus* were found in cave deposits; this in itself does not prove that these hominids were consistent cave dwellers. But the additional evidence of associated remains of stone and bone that seem to have been accumulated by these creatures—charred animal bones, collections of seeds, and what could be ancient hearths and charcoal—all point to *H. erectus* as having spent appreciable periods of time as a troglodyte (cave dweller) at Chou-k'ou-tien. The remains of Lan-t'ien, Trinil, Sangiran, and Modjokerto, as well as Ternifine and Olduvai, were all found in open sites, sometimes in stream gravels and clays, sometimes in river sandstones, in conglomerate and volcanic rocks, or in lake beds. These suggest that *H. erectus* also lived in open encampments along the banks of streams or on the shores of lakes; proximity to water was crucial to human survival. The remains of the Vértesszőllős individual showed that he occupied mud flats deposited along with minerals around springs in a tributary of the Danube River. These open presumed campsites revealed by excavation contain abundant stone implements and stone chips that seem to have resulted from manufacture, fractured and partly burnt bones of animals that could have been hunted for food, and traces of what appears to have been a hearth.

Thus, both Chou-k'ou-tien and Vértesszőllős have shown signs that Middle Pleistocene humans had a controlled mastery of fire. Indeed, outside a cave at Chou k'ou-tien, charcoal was found along with traces of a stone toolmaking industry in an open gully deposit that seems to be slightly older than the cave deposit itself (containing the bones of *H. erectus*). In this region of China, therefore, it has been observed that the earliest convincing indication of the use of fire by humans immediately precedes the earliest example of a cave being occupied by them. This supports the notion that successful cave dwelling by human creatures depended on their first having mastered fire.

Mastery of fire

It was not only cave dwelling that mastery of fire seems to have made possible. Able to keep warm, man was apparently able to move into colder climes; indeed, this factor may have speeded the migrations of ancient humans into the chilly, often glaciated regions of prehistoric Europe. Sooner or later, too, humans started cooking their food, thus reducing the work demanded of their teeth. This, in turn, may have played an important part in minimizing the evolutionary advantage of big teeth—cooked food needs far less cutting, tearing, and grinding than does raw food. This relaxation of the evolutionary selective pressure that favoured the survival of people with strong, big teeth may, in turn, have led directly to a diminution in the size of the teeth—one of the features that has come to distinguish *H. sapiens* from *H. erectus.*

Other signs of the culture of *Homo erectus* are the implements found in the same deposits as their bones. Chopping tools made from split pebbles characterize both the Chou-k'ou-tien and Vértesszőllős deposits; both are members of a so-called chopper–chopping-tool family of industries. At Ternifine, in northwestern Africa, *H. erectus* was found in association with totally different kinds of stone implements; these comprise bifacial hand axes and scrapers that have been characterized as representing what archaeologists call an early Acheulean industry. This is part of the great Acheulean hand-ax complex of human industry, remnants of which are found widely spread over large parts of Europe and Africa. An Acheulean industry is known also from Olduvai Gorge, as is a local, more ancient form of stone chopper manufacture known as the Oldowan industry; but the exact cultural associations of these stone tools with African *H. erectus* (as exemplified by OH 9) are uncertain.

Hence, *Homo erectus* has been found associated in some parts of the world with a chopper–chopping-tool tradition and in other places with an Acheulean bifaced hand-ax industrial complex. Numerous nonhominid animal bones occur also with the remains of *H. erectus;* sometimes these bones seem to have been broken deliberately or burned. From this evidence, it seems that *H. erectus* was a hunter. His bodily (including cerebral) endowment and manufactured equipment were so much superior to those of *Australopithecus* and *H. habilis* that it is highly probable that his food-collecting techniques, including hunting, also were better. Many scientists hold that *Australopithecus* and *H. habilis* were more scavengers than hunters, perhaps at best opportunistic hunters who seized their chance when a weak, young, sick, or aged animal crossed their paths. Indeed, many of the animal bones found in australopithecine deposits are of juvenile and aged creatures; and, although larger animal bones have been recovered from *H. habilis* deposits, these have exhibited tooth marks of nonhominid predators as well as cut marks. *H. erectus,* on the other hand, seems to have been a confirmed hunter, and his prey included animals of all age groups.

Diet

It can credibly be supposed that, as with present-day hunters such as the Kalahari San (Bushmen) and the Australian Aborigines, meat from the hunt formed only a part of the diet of *Homo erectus*. Other juicy morsels may have been furnished by snakes, birds and their eggs, locusts, scorpions, centipedes, tortoises, mice and other rodents, hedgehogs, fish, crustaceans, and numerous other edible forms of life. Even children could have caught many of these—as they do in the Kalahari today, before they are allowed to accompany the older men on the hunt. Vegetable food also must have played a big part in the diet of *H. erectus,* in the form of fleshy leaves, fruits, nuts, and roots. Accumulations of hackberry seeds, for example, were found in the Chou-k'ou-tien cave deposits.

There seems little doubt that *H. erectus* must have been omnivorous (as *H. sapiens* is today), for such a diet is the most opportunistic of all, and modern humans are the most opportunistic of all living primates. Emancipated from too narrow an environmental dependence, from too restricted a dietary regimen, humans have come to live off many diets, in many surroundings. *H. erectus* was probably one of the earliest of the great opportunists; and it is likely that his very opportunism endowed the species with evolutionary flexibility, with adaptability, and with a very plastic survival kit.

Another question that may be asked about *Homo erectus* culture is whether there is any evidence of ritual among these extinct people. There is no sign that they buried their dead; no complete burials have been found, no graves, no grave goods, no red ochre (a mineral used as a paint by later forms of hominids) on or around the bones. That cannibalism was practiced was once inferred from the Chou-k'ou-tien finds, but little evidence remains to support such a hypothesis.

EVOLUTIONARY IMPLICATIONS

Erectus as a possible direct ancestor of sapiens. A few workers have generally opposed the view that *H. erectus* was the direct ancestor of *H. sapiens.* L.S.B. Leakey argued energetically that *H. erectus* populations, particularly in Africa, overlap in time with more advanced *H. sapiens* and therefore cannot be ancestral to the latter. His arguments were based primarily on stratigraphic considerations and chronology, but some support for Leakey's point of view has come from the analysis of anatomical characteristics exhibited by the fossils. By emphasizing the distinction between "primitive" and "derived" traits in the reconstruction of relationships between species, several paleontologists have attempted to show that *H. erectus* does not make a suitable morphological ancestor for *H. sapiens.* Because the braincase is long and low, is thick-walled, and presents a strong supraorbital torus, they claim *H. erectus* to show derived (or specialized) characters not shared with more modern humans. At the same time, it is noted that *H. sapiens* does share some features—such as a rounded, lightly built cranium—with earlier hominids (*e.g.,* the smaller, or "gracile," *Australopithecus* species or *H. habilis*). It is clear that these paleontologists consider the early, gracile Plio-Pleistocene *H. habilis* to be more closely related to *H. sapiens* than is *H. erectus* (a conclusion with which Leakey would have agreed).

Leakey's argument against ancestry

These findings have not been widely accepted. Stratigraphic arguments concerning the overlap of (late) *H. erectus* and (early) *H. sapiens* are no longer compelling, and it is difficult to reject hypotheses of continuity between Middle and Late Pleistocene *Homo* populations on morphological grounds. When the effects of body size on scaling of the brain are taken into account, for example, it is apparent that gracile australopithecines (and *H. habilis*) have relatively rounded skulls because they are small-bodied creatures. General resemblances between these earlier hominids—with their more spherical, lightly constructed craniums—and *H. sapiens* therefore cannot be used to demonstrate a direct relationship. Instead, studies of size and scaling in human evolution have emphasized that representatives of *Homo* can be grouped into a reasonable ancestor–descendant sequence showing increases in body size through time. Despite having a heavier, more flattened braincase, *H. erectus* is not out of place in this sequence, and there is no reason to expect that hominid craniums should have remained subspherical in form as bodies grew larger in the Middle Pleistocene.

If this much is agreed, there is still uncertainty as to how and where populations of *H. erectus* evolved into early *H. sapiens.* This is a major question in the study of human evolution, and one that resists resolution even when hominid fossils from throughout the Old World are surveyed in detail. Several general hypotheses have been advanced, but there is still no firm consensus regarding models of gradual change (local evolution in different geographic regions) as opposed to scenarios of rapid evolution in one region followed by migration of the new populations into other areas.

Gradualistic views of the transition to Homo sapiens. A traditional view held by many paleontologists is that a species may be transformed gradually into a succeeding species in the same lineage. Such successive species in the evolutionary sequence are called chronospecies. The boundaries between chronospecies seem almost impossible to determine by any objective anatomic or functional criteria; what is left is the guesswork of "ruling off" at a moment in time to draw the boundary. Thus, competent authorities have seriously suggested that, in the last analysis, such a chronological boundary may have to be drawn arbitrarily between the last survivors of *H. erectus* and the earliest members of *H. sapiens.* The problem of defining the limits of chronospecies is not peculiar to *H. erectus;* it is one of the most vexing questions in paleontology.

Such gradual change with continuity between successive forms has been postulated particularly for North Africa (where *H. erectus* at Ternifine is seen as ancestral to later populations at Rabat, Temara, Jebel Irhoud, and elsewhere) and for Southeast Asia (where *H. erectus* at Sangiran may have progressed toward populations such as those at Ngandong and at Kow Swamp in Australia). Some workers have suggested that similar developments could have occurred in other parts of the world. Certainly the most thoroughly documented statement of this view has been provided by the American anthropologist Carleton S. Coon in *The Origin of Races* (1962). Coon departed from traditional gradualist theory when he postulated the

Coon's *The Origin of Races*

Painting by Zdenek Burian; reproduced with permission

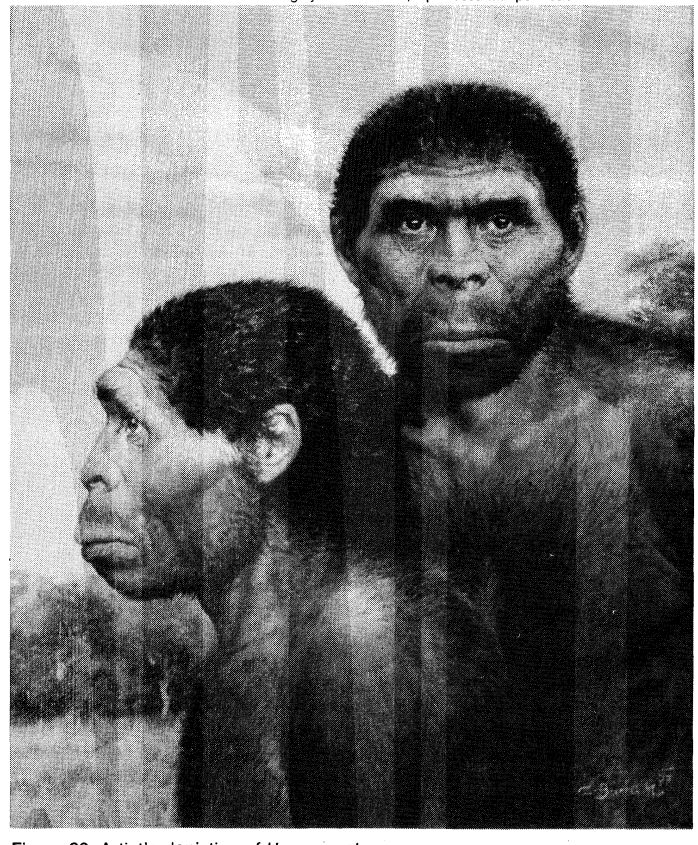


Figure 20: Artist's depiction of *Homo erectus.*

transition from *H. erectus* to *H. sapiens* as a crossing of a kind of evolutionary Rubicon—that is, a clear-cut boundary, or threshold, marked not only by bodily changes but also by striking cultural changes.

This supposed interrelation of cultural achievement and the shape and size of teeth, jaws, and brain is a theorized state of affairs with which many students of fossil hominids strongly disagree. Throughout the human fossil record, there are examples of dissociation between skull shape and size, on the one hand, and cultural achievement, on the other. A smaller-brained *H. erectus* from China seems to have been among the first humans to tame fire, while much bigger-brained people in other regions of the world, living later in time, have yielded no evidence that they knew how to handle fire. (Differences in inferred cultural activities between the African and Asian members of *H. erectus* have been emphasized above.)

The core of Coon's thesis is that *H. erectus* evolved into *H. sapiens* not once but five times, as each subspecies of *H. erectus,* living in its own territory, passed the postulated critical threshold. Since this part of Coon's theory depends on accepting his supposed *erectus–sapiens* threshold as correct, it is opposed by those who find the threshold concept at variance with the modern genetic theory of evolutionary change.

A corollary of this concept was Coon's suggestion that different subspecies (or "races") of man evolved at different rates and, hence, crossed the *erectus–sapiens* threshold at different times. In particular, Coon has suggested that "the step from the ancestral *H. erectus* to the modern *H. sapiens* was taken by Caucasoid man in Europe no less than 200,000 years before the same step was taken by Negro man in Africa." Not only was this part of Coon's reasoning based on a concept that is at odds with modern evolutionary theory, but some of the assumed "facts" on which it was based are highly problematic. He claimed that the races of today had already appeared before the *erectus–sapiens* transition. Apparently by allocating two very early fossil human skulls from Europe (Swanscombe from England and Steinheim from Germany) to the European, or Caucasoid, "race," although they lived 250,000 years ago, he concluded that the supposed *erectus–sapiens* transition had occurred among Europeans, or Caucasoids, as long as 250,000 years ago. Few if any paleoanthropologists would agree with Coon that these early skulls are to be recognized as those of members of the Caucasoid, or European, "racial" group, as generally understood.

Turning to Africa, Coon claimed that "Rhodesian man" (represented by parts of several skeletons found in Zambia [formerly Northern Rhodesia] in 1921) belonged to *H. erectus.* Coon, who thought that Rhodesian man lived only 30,000 or 40,000 years ago, thus argued that *H. erectus* was still present in Africa long after he had disappeared from the European scene. This argument hinges on controversial judgments of the species to which Rhodesian man is to be assigned and on dating of the fossil remains. In contrast to Coon, most paleoanthropologists regard Rhodesian man as a member of an African subspecies of *H. sapiens,* commonly dubbed *H. sapiens rhodesiensis.* Also, it is now thought that the bones are of later Middle Pleistocene age. The evidence for the late survival of *H. erectus* in Africa is questionable at best.

Alternative modes of species change. That phyletic gradualism marks the transition from *H. erectus* to *H. sapiens* is one interpretation of the fossil record, but the available evidence can also be read differently. Some workers who do not hold Coon's ideas of body change tied to cultural achievement have come to accept what can be termed a punctuational view of human evolution. In their view, species are seen as more stable entities, not simply as arbitrarily defined segments of a lineage. It has been suggested that species such as *H. erectus* may be expected to exhibit evolutionary stasis (*i.e.,* little or no morphological change) over long periods of time. The transition from one species to a descendant form may occur relatively rapidly in geologic terms and in a restricted geographic area rather than on a worldwide basis. Whether this may be a correct interpretation of the situation involving *H. erectus* and *H. sapiens* is unclear; that is, whether *H. sapiens* evolved from populations of more archaic *H. erectus* gradually, or rapidly during a short "pulse" of evolution late in the Middle Pleistocene, has not been settled, although a good deal of attention has been given to the problem.

Punctuational view of evolution

The continuation of the argument underlines the need for more fossils to establish the range of physical variation of *H. erectus* and for more discoveries in good archaeological contexts that permit increasingly precise dating. Additions to these two bodies of data may settle remaining questions and bring the unsolved problems of phylogeny and classification of *H. erectus* nearer to resolution.

(P.V.T./G.P.Ri.)

Homo sapiens

ORIGIN AND EARLY EVOLUTION

Homo sapiens, the hominid species that includes modern humans, emerged during the Pleistocene epoch. The date of the transition from the Pliocene to the Pleistocene is an arbitrary point in time, and geologists have not found it easy to agree on its definition; for years the generally accepted date was about 2.5 million years ago, but more recently a date of 1.6 million years ago has been adopted. Broadly speaking, it was marked by the gradual onset of a cooler climate in many parts of the world about 2.5 million years ago and by a general lowering of temperature that finally led to the great Ice Ages, during which, in modern temperate zones, ice caps and glaciers originating at high altitude and latitude spread out for considerable distances over lowlands. This process of glaciation was recurrent and extended throughout most of the Pleistocene epoch; it is now generally agreed that there were four main glacial periods, of varying duration and severity, separated by interglacial periods during which the climate became warmer and in some cases (even in Europe) almost subtropical. Evidence of the successive glaciations can be detected in the characteristic geologic deposits left by melting ice and also in the fossil remains of Arctic or subarctic animals and plants from the glaciated regions. By the determination of the rhythmic succession of glacial and interglacial phases during the Pleistocene, geologists have provided a time scale for inferring the relative antiquity of fossil hominids and the implements they left behind.

The Ice Ages

Even in those parts of the world where there was no actual glaciation, such as the equatorial regions, there was a succession of alternating rainy and dry periods, but there is no evidence that these pluvial and interpluvial phases were closely correlated with the glacial and interglacial periods in the Northern and Southern hemispheres. The recurrent glaciations were accompanied by considerable falls in sea level, which had a profound effect on the formation of river valleys and caused the opening of land bridges that permitted migration of hominid populations and, consequently, encouraged gene flow and mixture between such populations. Similarly, the rise in sea levels during the interglacials cut these routes and produced isolation of hominid populations. With a fall in sea level, the erosive power of the rivers increased, and they cut their valleys more deeply. With the rise in sea level during the interglacial periods, the rivers flowed more sluggishly and laid down such stratified deposits as gravel and sand over their alluvial plains. As a result, series of terraces were formed along the riverbanks, and it is in these terraces that some of the oldest remains of Paleolithic (Old Stone Age) humans and their stone implements have been found. The time relationship of the terraces to the glaciations has been worked out in some detail by geologists, and largely on this basis the relative antiquity of the fossils can be established.

Since the mid-20th century, new chemical and physical methods that have been developed have assisted greatly in establishing a chronology for *Homo sapiens* that enables paleontologists to set their materials in a proper phylogenetic context. Relative dating (the assessment of contemporaneity of a fossil and its layer) has been aided by the fluorine and uranium methods, while absolute dating (or age in years) has often established the sequence of fossils in the deposits.

The antiquity of Homo sapiens. It is a curious fact that,

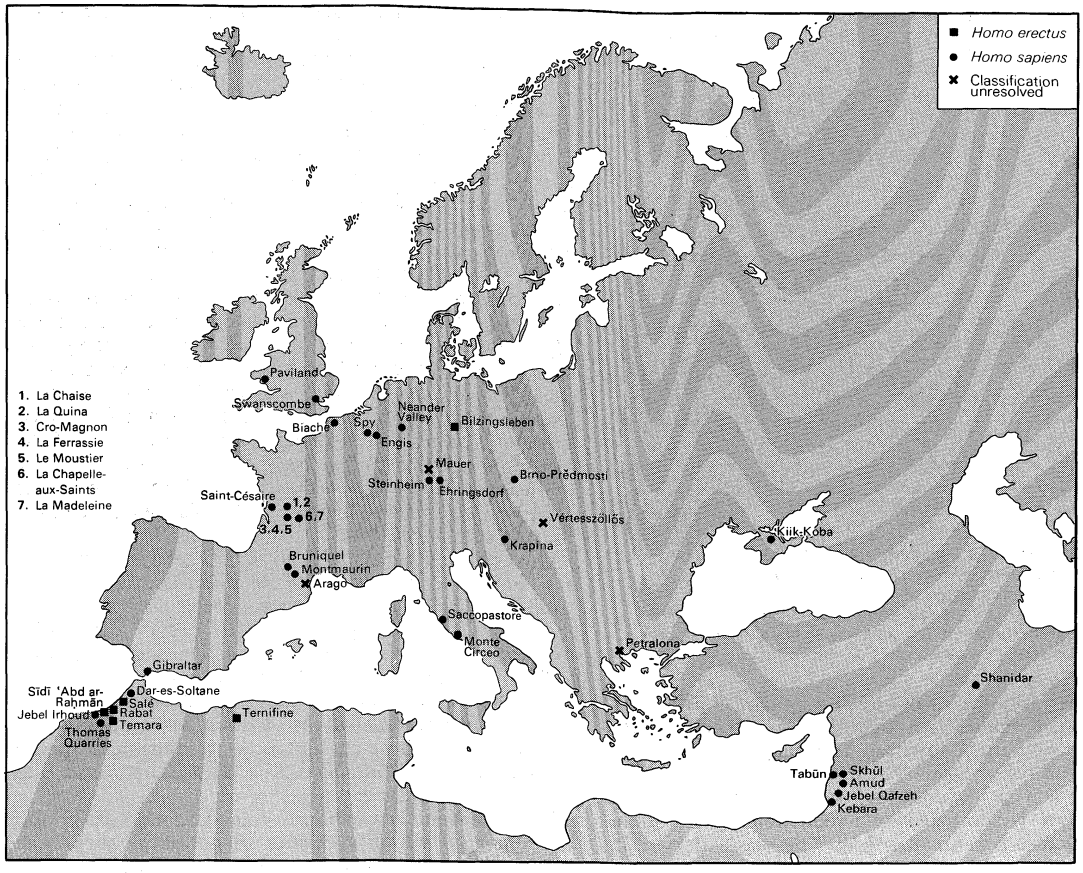


Figure 21: Major sites of hominid fossil finds in Europe, North Africa, and the Middle East.

Defining *H. sapiens*

although evidence for the evolution of man is extensive, direct fossil evidence of the earliest members of the species *Homo sapiens* is relatively scarce. The species *H. sapiens* (of which the modern human races comprise a number of different geographic varieties) may be defined in terms of the anatomic characters shared by its members. The definition for prehistoric representatives of the species must be limited to skeletal characters, the only remains to be found, and includes such features as a mean cranial capacity of about 1,350 cubic centimetres (82 cubic inches), an approximately vertical forehead, a rounded occipital (back) part of the skull with a relatively small area for the attachment of the neck musculature, jaws and teeth of reduced size, small canine teeth of spatulate form, the presence of a pointed or projecting chin, and limb bones adapted to a fully erect posture and gait. Any skeletal remains that conform to this pattern to an extent that precludes classification in other groups of higher primates must be assumed to belong to anatomically modern *H. sapiens.* In the past there was a tendency to create entirely new species of *Homo* on the basis of fragments of prehistoric human skeletons, even though the remains showed no significant differences from modern man. This tendency was prompted by the supposed antiquity of the remains or by a failure to realize how variable some features are even in modern man. The species of the genus *Homo* that immediately preceded *H. sapiens* was *H. erectus,* and it is most likely that sapient humans (*H. sapiens*) evolved from *H. erectus.*

Bodily structure of Homo sapiens. Fossil remains of early *Homo sapiens* are known from sites in Africa, the Middle East, and Europe, but later examples come from a wide range of sites in the Old World as a whole. It is of particular interest to look at the skulls of some of the early specimens, for it is in their functional morphology that the combination of features that results in attribution to *Homo sapiens* is often found.

Evolution of the human skull. The human skull is composed of both cranial and facial portions. The cranium consists of the skull vault and base, while the facial skeleton consists of the region of the eye sockets, nose, cheekbones, upper jaw (orbital and maxillary region), and the region of the lower jaw (mandibular region). During the evolution of the hominid skull from its apelike precursors, there are a number of general trends that can readily be discerned. The principal trends are the gradual increase in brain size (as measured by cranial capacity), the rounding of the cranial vault, and the gradual reduction of the size of the whole masticatory complex, including both the upper and lower jaws and the teeth. These trends lead to an overall change in skull shape and proportions, so that, while the vault expands, the "muzzle" tends to retract from a protruding (prognathic) form to a straighter-faced (orthognathic) appearance. At the same time, the whole skull tends to become lighter and more delicate in its structures. If this process is examined a little more closely, it can be at least partly explained in mechanical terms. A comparison of the skull of *H. erectus* and a modern *H. sapiens* skull illustrates the point (Figure 22).

The skull of *H. erectus* has a low skull vault, a sharply receding forehead, a lower cranial capacity (averaging less than 1,000 cubic centimetres), a large face with big teeth and jaws, and a jutting occipital region (the back of the head). Muscles in the occipital region relate principally to the balance of the head on the vertebral column (spine); those of the temporal region (at the sides of the head) relate to the working of the jaw apparatus. In a skull with a large face and a heavy jaw, the muscles of the occipital region, which balance the head on the neck (nuchal musculature), must be strong in order to counterbalance the heavy face and jaw. This results in marked occipital ridges in *H. erectus.* Similarly, large jaws and teeth demand well-developed muscles of mastication with strong attachments to the cranium. In order to dissipate the considerable forces produced by these muscles when the teeth meet during chewing, there is in *H. erectus* a stout maxilla (upper jaw) and a heavy supraorbital ridge, or torus (browridge), extending the whole width of the cranium, a ridge rendered even more necessary because *H. erectus* had a skull of lower cranial capacity and therefore lacked a well-developed forehead. Such skulls are found in the examples of *H. erectus* known from both China and Java.

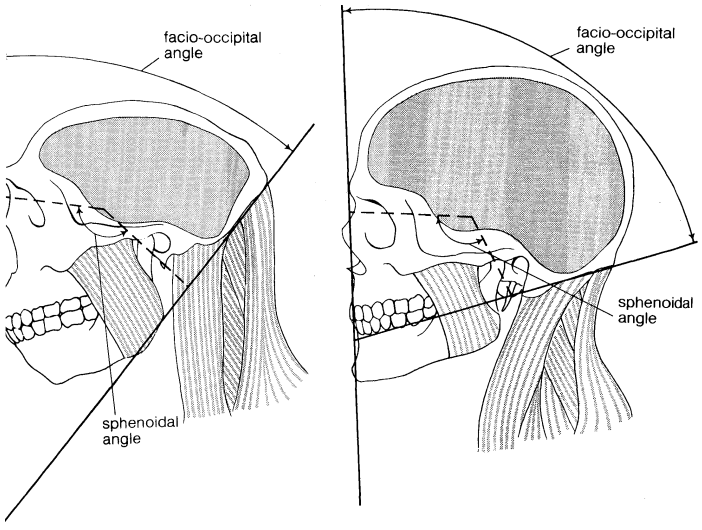


Figure 22: *Comparison of Homo erectus and Homo sapiens skulls, showing differences in cranial capacity, sphenoidal angle, and facio-occipital angle.*
(Left) Small brain, large muscles, and flattened skull vault of *H. erectus* skull. (Right) Large brain, small muscles, and rounded skull vault of *H. sapiens* skull.

The skull of *H. sapiens*

By comparison, the skulls of *H. sapiens* show an expanded cranial vault with a high maximum breadth and a well-developed vertical forehead, resulting from expansion of the frontal region of the brain, so that the supraorbital ridge is, as it were, "overgrown" and buried. At the same time, the face is shortened by a reduction in size of the jaws, which also bear smaller teeth. The masticatory muscles therefore need not be so large; and the forces that they produce are less, so that the need to neutralize them through the face is reduced. Thus, the facial structure can be more delicate, and the need for a heavy supraorbital ridge is removed. Finally, the lighter face does not require such powerful muscles behind the point of balance of the skull; as a result, the neck muscles can be reduced in size and their areas of attachment to the skull brought more underneath the back of the head.

The decrease in the size of the teeth of *H. sapiens* tends to leave the nose and point of the jaw as prominent features of the face. Thus, the presence of a mental protuberance (chin) is an obvious character of the jaw of *H. sapiens* and provides external support for the symphysial region (the point at which the two sides of the jaw have grown together).

Dental characteristics. The dental characteristics of *H. sapiens* revolve around the basic fact of reduction of the masticatory apparatus. Thus, the dentition as a whole shows tooth crowding (dento-alveolar disproportion), accompanied by smallness of the individual teeth and marked reduction in size of the third molar. In modern populations the third molars tend to be genetically unstable; *i.e.,* they are frequently absent or malpositioned (impacted wisdom teeth). Similarly, but less frequently, the lateral incisors (the cutting teeth on either side of the front teeth) may be absent. In other respects the dentition of *H. sapiens* is best contrasted with that of *H. erectus* in that it lacks some of the special features known from this species. For example, the teeth of *H. sapiens* are less likely to show secondary enamel wrinkling of the occlusal (chewing) surface of the teeth and less taurodontism (pulp space enlargement), although this is still well known from Neanderthals. The eruption sequence of teeth seems to be an unreliable criterion because variability is common in modern populations and may have been so in the past.

Postcranial skeleton. The form of the skeleton of the trunk and limbs of *Homo sapiens* (postcranial skeleton) is characterized by its adaptation for a fully upright posture and a striding bipedal gait. This remarkable locomotor capability is the final expression of an evolutionary process that has taken at least four million years to achieve, and so some aspects of the process are well known from earlier members of the genus *Homo* and also from the genus *Australopithecus.*

Bipedal posture and gait

In terms of posture, the bipedal vertebral column is held upright and shows two secondarily developed curves when viewed from the side, one in the lumbar region of the back (small of the back) and the other in the neck region. From the front the column should appear straight. These curves allow the weight to be evenly disposed about the line of gravity, which passes vertically through the second sacral vertebra (at the base of the spine) and behind the rotation centres of the two hip joints. This permits the pelvis to tip backward just beyond the vertical and rest upon a straplike ligament across the front of the hip joint, a sophisticated effort-saving mechanism that allows most of the muscles around the hip to relax so that the upright stance is an economical posture. Associated with this is the ability to lock the knees back, which also relaxes some surrounding muscles. To rise from the squatting or seated position requires considerable power of extension of the hip joints, and this is provided by the large buttock muscle (gluteus maximus) and a backward extension (posterior superior iliac spine) of the bony pelvic flange (blade of the ilium) for its attachment.

An alternating bipedal gait, to be fully efficient, must allow each leg to swing clear of the ground during walking; this is provided for by a pelvic-tilt mechanism that raises the side of the swinging leg. In addition, such a gait must avoid wild side-to-side movements of the centre of gravity, and this is achieved by inclining the thighbones toward the midline and thus bringing the feet closer together. Finally the bipedal adaptations of the modern human foot are such that both weight and force are transmitted to the ground through a propulsive system of short levers that permits a heel–toe stride.

The upper-limb adaptations to bipedalism are fewer and are concerned with the dynamic balance of the body while moving. Arm swinging is a normal part of bipedalism and compensates for the twist of the body toward the side opposite to the advancing foot. The selective advantages of bipedalism, in terms of the upper limb, are immense in that they free the hands for the carriage of infants, food, tools, or weapons, as well as permitting the development of the hands for a manipulative role such as toolmaking. Although hominids below the human level of evolutionary advance could make tools, the refinement and exploitation of tools demanded hands capable of both power and precision grips. The power grip involves primarily the inner, or ulnar, side of the hand and permits a firm grip on a branch, a rock, or hammer handle. The precision grip involves the outer, or radial, fingers and thumb, as in using a small stone for engraving, a small brush for painting, or a pen for writing (Figure 23). This requires the bringing together of the tips of the opposed thumb and the next two fingers in order to grip a small object, a grip that demands that the lengths of the index finger and thumb be proportionate and that the joint at the base of the thumb

Freeing the hands for tool use

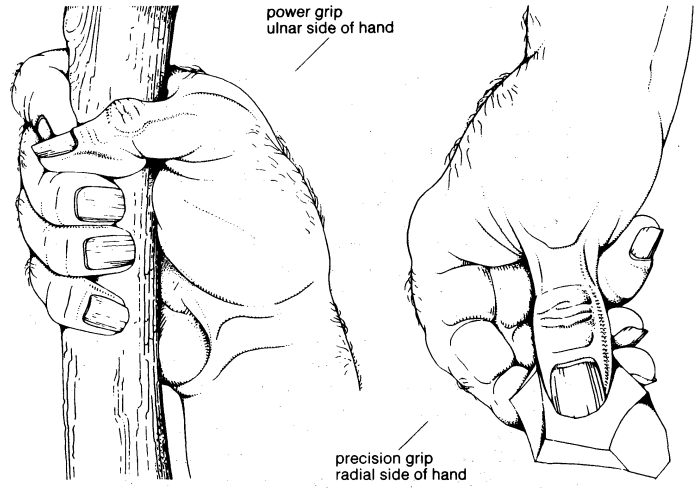


Figure 23: Hand grips.

be of a special saddle-shaped variety. It seems likely that the precision grip evolved later than the power grip and that its perfection may even have been a specialization in *Homo sapiens.* Only when human hands had evolved to this level, concomitantly with brain expansion, could manipulative skills give expression to the artistic impulse in terms of cave painting, bas-relief, and sculpture in the round, all of which are sophisticated behavioral correlates of a highly evolved individual, in terms of both locomotor and intellectual skills.

Fossil evidence. Having looked at the problems of the identification of human characteristics and also at the difficulties of interpreting the significance of these characteristics singly or in groups, it is appropriate to consider the early fossil materials that might provide reasonable candidates for attribution to *Homo sapiens.* These are likely to be found during the late Middle Pleistocene to Late Pleistocene (about 200,000 to about 15,000 years ago)—in archaeological terms, during the Paleolithic period, or Old Stone Age. Naturally, the quantity of fossil material recoverable tends to increase markedly as more recent periods are dealt with, so that the total number of human specimens amounts to some hundreds. On the other hand, toward the earlier end of this time range, the fossils available for study are far fewer in number. Advances in dating methods have eliminated a number of specimens formerly given an antiquity that they do not deserve, so that the picture has been somewhat clarified. A number of specimens remain that often in the past have been divided into the Neanderthals (or Neandertals) and the sapients (*i.e.,* members of *Homo sapiens sapiens,* the same subspecies as modern humans).

The problem of Neanderthal man. This group was originally recognized by a combination of cranial, dental, and postcranial features that were generally considered distinct enough to classify Neanderthals as a separate species. Thus *Homo neanderthalensis* was described on morphological grounds and regarded as a specialized group of the genus *Homo* that lived during the last glaciation (Würm glacial stage) in Eurasia. The sites from which examples of the Neanderthals were recovered have commonly produced tools of the Mousterian culture, a stone tool industry dating from about 90,000 to about 30,000 years ago. Quite suddenly these people disappeared from the fossil record, and various theories have been put forward to account for their disappearance. In addition to these "classic" Neanderthals—exemplified by remains from such sites as the Neander Valley (West Germany); La Chapelle-aux-Saints, Le Moustier, La Ferrassie, and La Quina (France); Gibraltar; Monte Circeo (Italy); Shanidar (Iraq); and Kiik-Koba (Soviet Union)—there are also remains from Krapina (Yugoslavia), Saccopastore (Italy), Ehringsdorf (East Germany), and Tabūn (on Mount Carmel), Amud, and Kebara (Israel).

Many have considered that the classic Neanderthals were a cold-adapted, specialized side branch from the human line that became isolated in Europe and then became extinct as the climate improved. Some Neanderthals are considered to have avoided this specialization and continued to give rise to the later modern sapient populations. This involves invoking the so-called catastrophic demise of classic Neanderthals. An alternative view placed classic Neanderthals in the modern human evolutionary line, their disappearance both anatomically and culturally being due to an absorption process involving some contribution of Neanderthal genes to the succeeding populations. Most modern systematists tend to include Neanderthals within the species *Homo sapiens* and only accord subspecific status to the combination of morphological characters that make up the anatomy of the classic Neanderthals of the Würm glaciation; they are thus named *Homo sapiens neanderthalensis* rather than *Homo neanderthalensis.* Some workers subsequently have advocated returning Neanderthal man to the status of a separate species.

Neanderthals' place in human evolution

In addition to the European and Middle Eastern evidence of Neanderthals, a number of specimens from Africa and Asia must be considered. Examples of this group could include Rhodesian man (Zambia, formerly Northern Rhodesia), Solo man (Java), and Saldanha man (South Africa), all from Late Pleistocene deposits (130,000 to 10,000 years ago). Their very presence shows that the whole problem of defining *Homo sapiens* must be considered in a wider context and cannot reasonably be considered in European terms alone. Clearly, the problem of the origin of *Homo sapiens* from his Middle Pleistocene forebears is complex; hence, it is valuable to examine in detail those specimens that come from the earliest well-dated sites, in order to try to discern the centre of sapient evolution (if such a centre exists) and to discover which are the earliest specimens that show incipient sapient characters. Bearing in mind the principle of mosaic evolution, which has been demonstrated as an important part of the evolutionary process, it is not reasonable to suppose that all of the features that have been mentioned as characteristic of *Homo sapiens* will appear together in these early specimens. Thus, a mixture of advanced and less advanced features may be expected in early forms, some perhaps relating to *Homo erectus,* to Neanderthals, or to modern humans as they are today.

The Omo hominids. In 1967 the Kenyan group of an international expedition to the Omo River in southern Ethiopia led by the paleoanthropologist Richard Leakey recovered large parts of two skulls and a large number of limb bones from two sites that have been dated to the East African late Middle Pleistocene or early Late Pleistocene (about 200,000 to 70,000 years ago). The most complete skull (Omo II) is long and narrow with a receding forehead, a rounded cranial vault, and a prominent occipital ridge, below which lies a flattened neck region. The bones of the skull vault are thick, and the maximum breadth of the skull is low in the temporal region, which is marked by the presence of large mastoid processes (bony projections below and behind the ears). The completeness of this skull has allowed accurate estimations of the cranial volume at about 1,430 cubic centimetres. Preliminary assessment of the taxonomic affinities of this skull has suggested that, while it shows a number of specialized features, some of which are reminiscent of the earlier *Homo erectus* skulls, it also has some advanced features, such as the rounding of the vault, the large mastoid processes, and the high cranial capacity. In view of this, it has been regarded as an early example of the African segment of evolving *Homo sapiens.*

Features of the Omo skulls

The other skull (Omo I) recovered from the same general area was accompanied by a large number of limb bones and is in many ways quite different from Omo II. Although the forehead still slopes from a prominent supraorbital (brow) ridge, the vault is very well rounded down to the occipital region. There is a prominent down-turned mastoid process. The vault of this skull is robust by modern standards, but it is less so than that of Omo II.

Viewed from the rear, the two skulls show striking differences—the Omo I skull is far more modern in outline. The limb bones of Omo I are rugged and well marked by muscle impressions, but in general they have few, if any, features that would distinguish them from the limb bones of *Homo sapiens.* A preliminary assessment of Omo I therefore suggests that, although it has some specialized features, they are far fewer than those of the Omo II skull; at the same time, its advanced features are perhaps more marked.

Perhaps the most striking feature of the whole Omo assemblage is once again the mixed character of each of the two principal specimens that are said to be of the same geologic age. The Omo II skull undeniably has features that can be paralleled in earlier forms of *Homo,* and at the same time the Omo I specimen seems to foreshadow the later morphology of modern *Homo sapiens,* particularly in the occipital region of the skull and in the limb bones. It is possible that the mosaic evolutionary process is seen here in microcosm. Not only is differential evolution occurring within individual representatives of contemporary populations but also between populations in one area of East Africa.

The Omo I remains were partly in situ in Member I of the Kibish Formation, a layer of undeformed Pleistocene lake and river sediments to the north of Lake Turkana (Lake Rudolf). The Kibish Formation consists of four members

(major strata), the lowest of which (Member I) is more than 40 metres (130 feet) thick and has seven stratigraphic subunits. It was from Member I that the Omo I skull was recovered, while the Omo II skull has been referred to the same level in an analogous stratigraphic sequence.

The dating of Omo

The relative dating of the site has been established by means of the fluorine, nitrogen, and uranium tests, and the fauna recovered is not inconsistent with the late Middle Pleistocene date (*c.* 200,000–100,000 years ago), but it cannot be positively dated to this time level. The stone flakes are also undiagnostic in terms of relative archaeological dating. Direct chronometric dating has been attempted on material from the deposit, making use of both the carbon-14 and the uranium–thorium radiometric methods. The carbon-14 test has shown that the upper part of Member III, Member II, and Member I are all too old to be dated by this method and, therefore, older than 40,000–35,000 years. Member I of the Kibish Formation has been dated to 130,000 years ago by the uranium–thorium method, and thus both Omo I and Omo II are regarded as being of this age because their sites are stratigraphically equivalent.

Ngaloba man. In 1976 a skull was found in the Ngaloba beds at the Laetoli site in Tanzania. These beds consist of deposits of sandstone and clay stone that are preserved in patches; these patches are made up of redeposited detritus that has been eroded from the underlying Ndolanya and Laetoli beds. In the upper Ngaloba beds, stone tools have been recovered that have been attributed to the African Middle Stone Age cultural complex, and they resemble tools recovered from the upper Ndutu beds at nearby Olduvai Gorge. The dating of the Ngaloba beds at Laetoli is based on correlation with a marker tuff in the lower unit of the Ndutu beds at Olduvai, and its age is estimated at about 120,000 years.

The Laetoli skull (LH 18) is nearly complete, including most of the cranium and much of the face. The forehead is recessed, the occiput is rounded, and the mastoid process is small. The supraorbital torus is divided. The skull clearly belongs to an archaic form of *H. sapiens;* its morphology is modern, although archaic features are retained. The skull differs considerably from those of *H. erectus* or of Neanderthal hominids. Its greatest resemblance is with the craniums of other archaic *H. sapiens* such as Kabwe (Rhodesian man) and Omo I. If the date of this skull is accepted it is one of the earliest examples of *H. sapiens* that is known from Africa and adds to the view that this species of human arose there.

Border Cave man. The Border Cave site lies on the boundary between the black state of KwaZulu in South Africa and Swaziland. It has been known for many years and has from time to time produced human remains and artifacts. The various levels in the cave have been dated by correlation, but more recent carbon-14 dating of the Middle Stone Age layers showed them to be more than 48,000 years old. Newer data has shown that the hominid finds 1, 2, and 3 from the Border Cave are between 110,000 and 90,000 years old. These estimates are based on the correlation of climatic and cultural evidence derived from other African sites.

The remains consist of fragments of a skull (Border Cave 1), a mandible (Border Cave 2), and parts of an infant's skeleton (Border Cave 3). The features of the skull are very close to those of anatomically modern man and are without doubt to be attributed to *H. sapiens.* If the dating is correct, Border Cave man is another example of early *H. sapiens* from the African continent, this time from the south rather than the east.

Jebel Qafzeh. Jebel Qafzeh, a cave site in Israel not far from Nazareth, has also been known for many years, and the remains of about a dozen individuals have been recovered from it. The cave is large, and 24 layers have been identified that contain human remains and stone tools of the Levalloiso-Mousterian type. Animal remains are extensive, including those of horses, rhinoceroses, fallow deer, wild oxen, and gazelles. The site has been dated on the basis of the fauna and by a technique called amino-acid racemization, as well as by archaeological appraisal. The earlier evidence suggested an age of between 78,000 and 39,000 years, but subsequent evidence has suggested that the skeletons should be dated to 115,000 to 90,000 years ago.

The most completely reconstructed skulls show that they have no resemblance to either *H. erectus* or to Neanderthals but rather a clear affinity to *H. sapiens.* The cranial vaults are high and well rounded with no exaggerated browridges, and one skull, Qafzeh 6, has a cranial capacity of about 1,570 cubic centimetres. The limb bones are similar to those of modern humans. Once again, if the dating is correct, these hominids represent early examples of *H. sapiens* from the Middle East, thus linking the African populations with those of Europe.

Vértesszőllős man. In 1965 some fossil remains were recovered from a site in Hungary at the foot of the Gerecse Mountains. In a quarry cut into the fourth terrace of the Danube River system, a number of occupation layers were recognized. The third of these layers contained some human remains as well as a tool culture and some fossil mammalian bones. The site is known as Vértesszőllős, and the first finds were some fragments of milk teeth from the mandible of a child (Vértesszőllős I). Rather more important was the second find (Vértesszőllős II), a fine adult occipital bone that was broken into two pieces. The bone is stoutly constructed, having thick walls in the ear region, whereas the floor of the depression for each cerebellar hemisphere (at the base of the skull) is relatively thin. When seen in profile, the bone is divided into two parts: an upper, curved occipital portion and a lower, flattened neck portion. The flattened area is incomplete and does not include the margin of the foramen magnum (the opening at the base of the skull through which the spinal cord passes as it merges with the medulla oblongata), the borders of which are broken away. The attachments of the neck muscles are well marked on the outer surface of the flattened area, which also has signs of an occipitomastoid crest (a crest of the occipital bone near the mastoid process), a primitive feature. The occipital ridge, or torus, that divides the upper, curved portion from the lower, flattened area is prominent and continuous across the bone; indeed, it is so prominent that in its central portion it is even undercut. On the inside of the skull, the cerebellar fossae (depressions in the skull where the lobes of the cerebellum lie) are rather small compared to the impressions above for the occipital lobes of the brain. The internal occipital protuberance lies well below the external occipital protuberance (inion), and there are distinct impressions for the venous sinuses (marks where blood vessels lie against the skull). The cranial volume of this individual has been estimated at about 1,400 cubic centimetres.

Features of cranial remains

The remains have been attributed to a male less than 30 years old. The thickness and the breadth of the bone and the undivided occipital ridge are relatively primitive features, but the height and curvatures of the upper segment are considered modern; the brain configuration is primitive in spite of its large size. The morphological comparisons and the statistical analysis of this specimen suggest that, although this population took its origin with *H. erectus,* it had differentiated from this group and perhaps ought to be classified as at the beginning of a progressive evolutionary line.

Perhaps the principal significance of this specimen, apart from its structure, is that it is dated to a warm phase within the second (Mindel) glaciation, 500,000 to 400,000 years ago. The specimen has been assigned to an unusual new category *Homo* (*erectus seu sapiens*) *palaeohungaricus;* whatever the precise taxonomic meaning of this name, it is clear that it is regarded as an intermediate form, the mixed *H. sapiens* and *H. erectus* features of which are the result of mosaic evolution.

Petralona man. In 1960 a cranium was recovered from a cave in Thessaloníki *nomós* (department) in eastern Greece. The cave is part of an extensive system in calcareous Mesozoic deposits containing a large quantity of fossil mammalian bones, including horse, cave bear, cave lion, and others. Three superimposed faunas have been identified, and the skull has been linked to the oldest of the three. This suggests that the skull is older than the later Middle Pleistocene. The date of the skull is disputed, but

Age of the Petralona skull

the general consensus is that it may be as old as 400,000 years. This still makes the skull one of the oldest human finds from Europe. The question of its relationships is also unclear. While the skull seems more advanced than *H. erectus,* it has some *H. erectus* features; on the other hand, it can also be distinguished from both Neanderthals and from modern *H. sapiens,* while possessing some features that are more advanced than *H. erectus.*

Tautavel man. The Arago cave near Tautavel, in southern France, has been excavated since 1964 and has yielded numerous fossils of mammals as well as stone tools attributed to the Tayacian industry. The dating of the site has been a matter of some doubt; at first it was given as about 200,000 years ago and subsequently as early as 400,000 years ago. In either case this is an early date for human remains in Europe. The Arago skull was found on a living floor (*i.e.,* layer of human occupation) and consists of a partially deformed face and part of a vault that has been reconstructed. The result is equivocal: in some ways the skull resembles *H. erectus* and in other ways Neanderthals. On this basis it has been termed a "pre-Neanderthal." There is no doubt, however, that it shows not only archaic *H. erectus* features but also advanced characteristics which may be related to adaptation to cold in an early and primitive example of *H. sapiens.*

Swanscombe man. A more widely known example of early *H. sapiens* is derived from the Thames Middle Gravels at Swanscombe in North Kent, Eng., which are gravels attributed to the Mindel–Riss interglacial stage, 400,000 to 200,000 years ago. Here, at a site well known for its Paleolithic tools, two parts of a human skull were recovered in 1935 and 1936. About 20 years later, a third piece was recovered that fits with the other two pieces to form the back half of a cranial vault. All three bones are virtually complete, and it is believed that they belonged to a young adult. Generally speaking, the bones are modern; the bones of the skull vault are rather thick, however, and the skull is broad at the back. The occipital bone shows no sign of the heavy occipital torus, or "chignon," characteristic of Neanderthals, and the occipital ridge is modest, although a little more prominent at its ends than in the middle. This bone is rounded both above and below the ridge, in contrast with the Vértesszőllős occipital. The position of the foramen magnum and the joints for the vertebral column are modern, so that the posture of the head does not appear to differ significantly from that of modern man. In the original report on these bones, it was stated that by measurements the Swanscombe skull could not be distinguished from that of *H. sapiens.* Despite this, several authorities have been less convinced by the sapient features of Swanscombe man and have emphasized the Neanderthal features of the bones. A more recent reappraisal of the Swanscombe material, making use of multivariate statistics, suggested that it is necessary to emphasize the interrelatedness of all the early forms of *Homo* and that it is no longer possible to maintain specific status for each form. It is preferred, rather, to regard them all as a spectrum of varieties within one species. It was suggested further that Swanscombe should be regarded as belonging to a "Neanderthaloid Intermediate" group that contains the Ehringsdorf, Skhūl V, Krapina, and Steinheim specimens.

Classifying Swanscombe man

Once again, as with Vértesszőllős, there is a mixture of features that, at this somewhat later date in Europe, combines sapient and Neanderthal characteristics in keeping with the concept of mosaic evolution.

Steinheim man. Another skull of approximately the same age as the Swanscombe skull is known from the Steinheim site, near Stuttgart, W.Ger. This skull was recovered in 1933 from a gravel pit cut into Pleistocene deposits that have been dated to the Mindel–Riss interglacial stage. This skull is more complete than either of those mentioned above but suffers from some distortion, perhaps due to the pressure of the deposits. It consists of the cranium and the right side of the face of a young adult; much of the base of the skull is missing, but there is a good deal of the palate and also some teeth present, including a premolar and all the molars of both sides. The face is straight, with little projection of the upper jaw, while the vault of the skull appears to be long and narrow but fairly well rounded in profile, and there is only moderate frontal flattening. The occipital region is marked by a very low occipital ridge, and it does not have a Neanderthal "bun." The frontal region has a pronounced but divided browridge. The cranial volume has been estimated variously between 1,070 and 1,175 cubic centimetres, but the distortion precludes accurate measurement.

The teeth are of particular importance because they are perhaps the earliest sapient teeth known. The premolar crown is symmetrical in shape but has a large outer and a smaller inner cusp. The molars decrease in size from front to back, but the third molar is markedly smaller than the other two. All of the teeth have some degree of taurodontism (enlarged pulp cavities).

Again, there is a mixture of primitive and advanced features in this skull, the principal resemblances of which are, with the Swanscombe skull, described above. The contour of the occipital region and the rounded vault are advanced sapient features, whereas the broad nasal opening and the browridge recall the Neanderthal shape. The teeth are small and sapient in both form and molar size, while the marked reduction of the size of the third molars is a distinctly modern feature. The enlarged pulp cavities are sometimes regarded as primitive since they are widely known from both *Homo erectus* and Neanderthal teeth. However, this trait is still frequently observed in the teeth of modern populations.

Taxonomic affiliations

Earlier opinions have suggested that Steinheim man represented an intermediate stage between *H. erectus* and the later forms of *Homo,* thus being ancestral to both Neanderthal and modern man. Another view does not accept the Steinheim fossil as being ancestral to the classic Neanderthal and modern man, suggesting instead that it represents a stage on a separate but parallel line leading to modern man. Both views now appear to be too simplistic and not in harmony with the findings of modern population genetics; the pattern of mixed characters that is becoming apparent follows the general trend of cranial enlargement and rounding, accompanied both by facial and dental reduction.

Behavioral inferences. While bones and teeth provide the most direct evidence of the form and locomotor capabilities of fossil man, there are other sources of information from which to obtain some concept of both his surroundings and his activities. This information comes from the careful excavation of the geologic deposits from which the fossil remains are taken. Modern methods in archaeology make use of an increasing number of scientific techniques, and through these studies it is sometimes possible to reconstruct both the environmental situation and the behavioral activities of fossil humans at a given site and time period.

Cultural remains preserved in the fossil record. In considering early *Homo sapiens* from a cultural standpoint, there are a few general considerations to be taken into account. The evidence that can be found must be inherently capable of preservation; thus, the types of materials most likely to be found are stone tools or stone structures and often fossilized animal bones and teeth. On the other hand, wood and plant remains are much rarer because such materials are more subject to decay. It must not be assumed, therefore, that the absence of wooden materials in the fossil record implies that they were not used by early man. Materials not normally preserved may be present in some situations and provide valuable clues; hence, fossil pollens, dealt with by the science of palynology, can give enormous insight into the vegetational structure and thus the climate of the time. Other general considerations relating to a fossil site include the likelihood of its location being close to a source of water for both humans and the animals that they may have hunted. Similarly, it is likely that sources of stone for the manufacture of tools will be found near living sites. Given these basic requirements, the living sites of early humans can clearly be in temporary structures in the open, in caves, or in rock shelters. In many circumstances it is likely early humans had little choice about where to shelter, but under glacial conditions the ability to find, acquire, and defend a cave or shelter

could have been crucial and might have overridden other considerations. During interglacial periods, when warmer conditions prevailed in summer in Europe and possibly all the year round nearer the equator, a temporary open-air site with a makeshift windbreak or skin tent may have sufficed.

Homo sapiens' early distribution

The known occupation sites of early *Homo sapiens* are so few that it is unlikely that any valid distribution pattern can be described at this stage. The principal early sites are from Europe and Africa, but they vary in age to such an extent that it would be unwise to draw conclusions as to possible migrations or to population density at any given time period.

Stone tools associated with Homo sapiens. Having considered the fossil remains of early *Homo sapiens* and the environmental and cultural situation of the group by means of the analysis of remains associated with the fossils, it is possible to look further by considering other sites that have provided tools but no human remains. Sites of this kind are very numerous and it is easy, but unwise, to speculate as to the kind of individuals that occupied them on the theory that human types and tool types will invariably go hand in hand. Experience has shown that such a correspondence cannot be safely inferred in most cases because overlap of both differing human populations and differing tool cultures is known to exist.

In general terms, the earliest known stone tool industry from the Early Pleistocene (about one million years ago) seems to have split into two basically parallel cultural systems, the hand ax and the flake tool industries (Figure 24). Within each of these types there is evidence of late diversification, so that the variety of tools increases; each presumably was used for more specific purposes.

Hand ax industries

The distribution of tool sites shows some general features, such as the restriction of hand ax industries to Africa, western Europe, Arabia, and the Indian subcontinent, whereas the flake industries overlap with the hand ax industries in western Europe but extend through the Balkans and into Southeast Asia. It seems clear that in Africa hand axes were developed from chopping tools, such as those found at Olduvai Gorge in Tanzania, by the gradual extension of the flaking process around the edges until the flaked areas met on the other side of the stone. The earlier stages of this process have been shown at Olduvai, but it is uncertain whether all hand ax cultures derived from a single African dispersal centre. The early Oldowan chopper tools led eventually to the principal group of hand axes known as the Acheulean industry and it seems very likely that early Acheulean hand axes were made by *Homo erectus.* This does not mean, however, that the later products of that industry must also be attributed to *H. erectus,* as has been shown by the fine Acheulean tools recovered from Swanscombe and many other sites of much later date.

Flake tool industries

The parallel, but not mutually exclusive, flake industries seem to have originated in the large chopper tool industries, such as those made by *H. erectus* at Peking (Choukoutienian industry). Similar early flake industries are known from Burma (Anyathian), Malaysia (Tampanian), and from the Indian Punjab (Soan). The early Buda industry of Vértesszőllős had no hand axes but many choppers and flake tools yet associated with *Homo sapiens.* The principal early development of the flake industry was the Clactonian industry, which is known from a number of European sites of the Mindel–Riss interglacial stage (400,000 to 200,000 years ago). In this industry the choppers and chopping tools were made from large flint nodules by striking off coarse but useful flakes. In its later and more evolved form, this tool industry has been termed Tayacian (it dates from *c.* 200,000 BC) and is known from sites in the Dordogne region of southwestern France. Beyond this, the flake and hand ax industries seem to merge into the Mousterian phase, while the more advanced Levallois technique leads on the Solutrean industry. (These are all stone tool industries of the Upper Paleolithic in Europe—*c.* 80,000–15,000 years ago.) All of the latter industries are well associated with *Homo sapiens* of the Late Pleistocene (after *c.* 100,000 years ago). (M.H.D.)

From (top) C. Ovey, *The Swanscombe Skull,* Royal Anthropological Institute of Great Britain and Ireland; (centre, bottom) K. Oakley, *Man the Tool-Maker;* used with permission of the Trustees of the British Museum (Natural History)

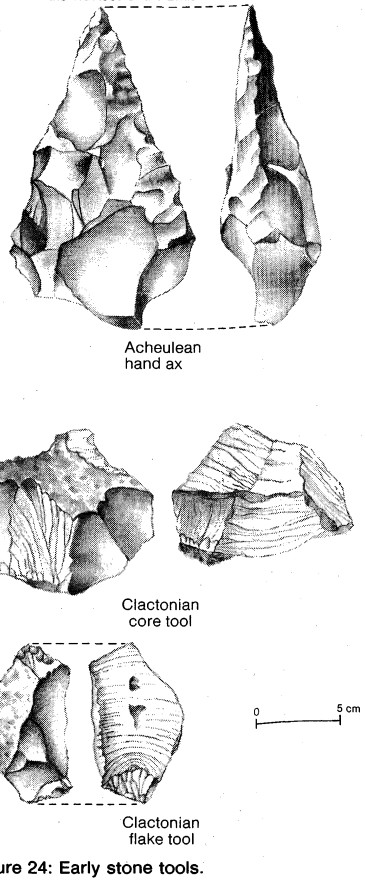


Figure 24: Early stone tools.

NEANDERTHALS

The human populations referred to as the Neanderthals inhabited much of Europe and western Asia during the time period of the earlier Late Pleistocene. They were the most recent archaic humans and represent the immediate predecessors of early modern humans. They lived in the region from Atlantic Europe eastward to Central Asia and from the latitude of Belgium southward to the Mediterranean and the Levant. The earliest of the Neanderthals emerged from more archaic humans between 200,000 and 100,000 years ago, and they were replaced by early modern humans between 50,000 and 30,000 years ago across the same region. Similar human populations lived at the same time in eastern Asia and Africa. As one of the more recent groups of archaic humanity, one that lived in a region of abundant limestone caves (which preserve bones well) and where there has been a long history of prehistoric research, the Neanderthals are better known than any other archaic human group. They have consequently become the archetypal cavemen, a role they play only reluctantly.

Early discoveries

Fossil finds. The name Neanderthal (or Neandertal) derives from the Neander Valley near Düsseldorf, W.Ger. In a cave in the valley in 1856 quarrymen unearthed portions of a human skeleton; 16 pieces of the skeleton were rescued and described shortly thereafter. Immediately, there was disagreement as to whether the bones represented an archaic and extinct human form or an abnormal modern human. The former view was shown to be correct in 1886, when two Neanderthal skeletons were discovered in a cave at Spy, Belg., associated with Middle Paleolithic stone tools and an extinct subarctic fauna. However, the view that the Neanderthals were somehow pathological still survives, despite abundant evidence to the contrary.

Shortly after the Spy discovery, up to about 1910, a series of Neanderthal skeletons were discovered in western and central Europe. Using those skeletons as a basis, scholars reconstructed the Neanderthals as semihuman, lacking a full upright posture and being somewhat less intelligent than modern humans. According to that view the

Table 3: Cultural Correlations in the Paleolithic Period

geologic period	years ago (000)	hand ax industry	flake industry	sapient site
Würm III glaciation		Magdalenian		Neanderthal and later sapient sites
Paudorf interstadial		Solutrean		
Würm II glaciation		Aurignacian		
Gottweiger interstadial		Perigordian		
Würm I glaciation		Mousterian	Tayacian	
Third interglacial	100			
(Riss–Würm)	130		Levalloisian	Omo
Riss II glaciation		Acheulean		
Inter-Riss interstadial				
Riss I glaciation			Clactonian	Steinheim Swanscombe
	250			
Second interglacial (Mindel–Riss)				
Mindel II glaciation			Buda	
Inter-Mindel interstadial	350?	Abbevillian		Vértesszőllős
Mindel I glaciation				
First interglacial (Günz–Mindel)				

Neanderthals were intermediate between modern humans and the apes (no older human forms were then generally recognized) but also too divergent to be the ancestors of modern humans. Only after World War II were the errors in the old Neanderthal postural reconstruction recognized, and the Neanderthals have since come to be viewed as quite close to modern humans evolutionarily. This view has been reflected in the frequent inclusion of the Neanderthals within the species *Homo sapiens,* usually as a distinct subspecies (*H. sapiens neanderthalensis*); some scholars maintain them in a different but closely related species, *H. neanderthalensis.* In the meantime a number of Neanderthal skeletons have been found in caves and shelters across Europe, the Middle East, and eastward to Uzbekistan in Central Asia, providing abundant skeletal remains and associated archaeological material for understanding these prehistoric humans. The Neanderthals are now known from several hundred individuals, represented by remains varying from isolated teeth to virtually complete skeletons.

Neanderthal origins and anatomy. The fossil evidence for the few hundred thousand years leading up to the time of the Neanderthals shows a gradual change from a *Homo erectus* form to one approaching the Neanderthals. Particularly in western Europe, the evidence shows a gradual decrease in the size and frequency of the anatomic characteristics of *H. erectus* and a gradual increase in features suggestive of the Neanderthals. From this a gradual emergence of the Neanderthals from earlier regional populations of archaic humans can be inferred, especially in western Europe and probably across their entire geographic range.

By courtesy of the Musee de l'Homme, Paris

Figure 25: Side view of the La Ferrassie 1 skull, that of an adult male Neanderthal from western Europe.

The changes between Neanderthal ancestors and the Neanderthals highlight their characteristics. Brain size gradually increased to reach modern human volumes relative to body mass, even if Neanderthal brains and braincases tended to be somewhat longer and lower than those of modern humans. Neanderthal faces remained large and especially long, similar to those of their ancestors and retaining browridges, a projecting dentition and nose, and the absence of a full chin. Their premolars and molars were reduced to the size of early modern humans, and their chewing muscles and cheek regions were reduced accordingly; yet their incisors and canines remained large, like those of their ancestors, indicating continued use as a vise or third hand.

Characteristics

The postcranial skeletons of the Neanderthals changed little from those of their ancestors. They retained broad shoulders, extremely muscular upper limbs, large chests, strong and fatigue-resistant legs, and broad, strong feet. There is nothing in their limb anatomies to indicate less dexterity or an inability to walk bipedally. The details of their hand bones, however, suggest greater emphasis on power than precision grips. Furthermore, details of their pelvises, legs, and toes suggest that they engaged in more irregular, lateral movement during locomotion than do most modern humans. All of these features, however, appear to have been inherited and maintained from their ancestors.

© Erik Trinkaus

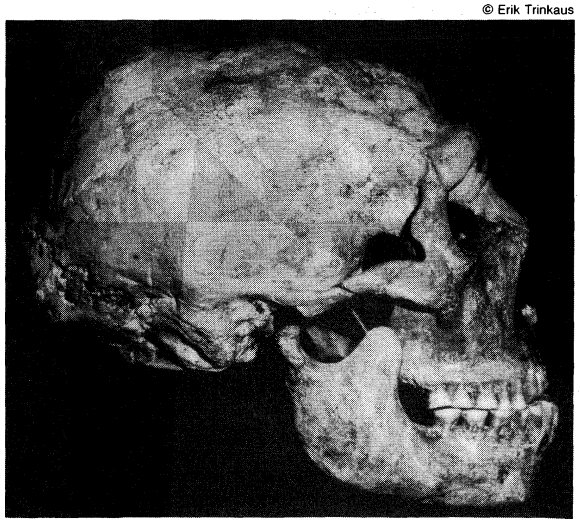


Figure 26: Side view of the Shanidar 1 Neanderthal skull, found at Shanidar Cave, Iraq.

The Neanderthals differed from their East Asian and African relatives (other late archaic humans) primarily in their retention of large incisors and canines and long faces to support those teeth. In Africa and East Asia front teeth and faces became smaller; at the same time, in both these and Neanderthal populations, the size of premolars and molars and facial massiveness were diminishing.

The fate of the Neanderthals. The evolutionary fate of the Neanderthals is closely related to the origins of modern humans. Over the years, the Neanderthals have been portrayed as everything from an evolutionary dead end to the direct lineal ancestors of modern European and western Asian populations. The evidence now indicates that modern humans first emerged in sub-Saharan Africa sometime prior to 50,000 years ago. Subsequently they spread northward, absorbing and occasionally displacing (through competition, not confrontation) local late archaic human populations. As a result the Middle Eastern, Central Asian, and central European Neanderthals were absorbed into those spreading modern human populations, contributing genetically to the subsequent early modern human populations across those regions. In western Europe—a culde-sac where the transition to modern humans took place relatively late—it is probable that the local Neanderthal populations died out largely without issue.

Emergence of modern humans

The anatomic changes between the Neanderthals and early modern humans involved largely a loss of the massiveness characteristic of all archaic humans. Limbs became more gracile, although they were still very muscular by modern human standards. The hand anatomy shifted to emphasize precision grips, and the frequency of lateral movement during walking was reduced. Front teeth became smaller and faces shortened, producing full chins and brows without ridges. Braincases became more elevated and rounded, but brains became no larger, nor, as far as is known, did humans become smarter.

Neanderthal behaviour. The behavioral patterns of the Neanderthals can be inferred from their anatomy in combination with the debris they left behind, their archaeological record. From their fossil remains and the hundreds of sites created by them—in cave entrances, rock shelters and the open air—an accurate view of their lifeways can be put together.

The Neanderthals appear to have lived in relatively small groups, moving frequently on the landscape but reusing the same locations often. This is indicated by the small sizes of their sites and by the considerable depth of debris at a number of sites. The materials left behind show only minor variations among sites, suggesting that there was little planned differential use of the landscape, one site serving as well as another for most purposes. In fact, combined with the evidence from their leg bones for high levels of irregular movement over the terrain, their sites suggest opportunistic use of resources and little forethought in their daily foraging activities.

Tool kits

Their tool kits, those subsumed by the Middle Paleolithic, or Mousterian, technological complex, included carefully made chipped stone tools or broad flakes and simple spears made of wood. Although much of their stone technology was simple and crude, they occasionally made high-quality stone tools by first preparing the block of raw material so as to strike off symmetrical and relatively uniform stone flakes. They rarely used bone as a raw material, however, despite its abundance in their sites as kitchen debris, and few of their tools were hafted. The predominance of hand-held thick flakes in their tool kits matches the strength of their arms and hands; such tools would have required great strength to carry out the same tasks that modern humans accomplish with more mechanically efficient implements and less strength. It also fits with their tendency to use their front teeth as a vise, augmenting their hands and tools.

Information about the Neanderthal diet—all of their food was gathered from the landscape—consists mostly of the animal bones that they left behind. There is rare evidence that they ate nuts, tubers, and other plant foods when available. The animal bones they abandoned indicate that they were able to hunt small and medium-sized animals (goats and small deer) but were able to obtain food from larger animals (*e.g.*, elk, horses, and cattle) only by scavenging from large carnivore kills. The species that were exploited closely reflected what was available in the surrounding countryside. Any use of small mammals, fish, birds, or shellfish as food must have been rare. There is simply no evidence for any systematic harvesting of wild plant or animal resources, a characteristic of modern hunter-gatherers in similar environments.

© Erik Trinkaus

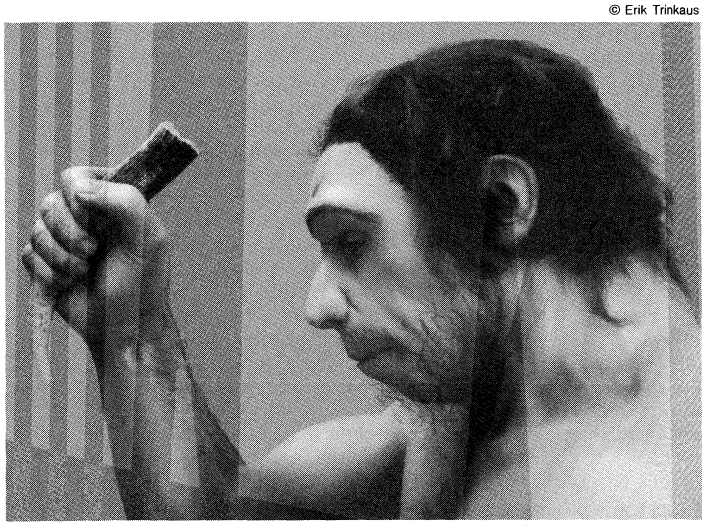


Figure 27: Reconstruction of the appearance of a Neanderthal man. Shown are the head and shoulders of a complete statue.

At the same time, the Neanderthals were the first human group to survive in northern latitudes during the cold (glacial) phases of the Pleistocene. They had domesticated fire, as indicated by concentrations of charcoal and reddened earth in their sites. Yet, their hearths were simple and shallow and must have cooled off quickly, giving little warmth through the night. Not surprisingly, they exhibit anatomic adaptations to cold, especially in Europe, such as large body cores and relatively short limbs, which maximize heat production and minimize heat loss.

The Neanderthals, despite their archaic anatomy and their less efficient foraging systems (compared to those of modern human hunter-gatherers), exhibited some uniquely modern features. They were the first humans to intentionally bury their dead, usually in simple graves. This indicates sufficiently elaborate social systems to make some kind of formal disposal of the dead desirable. They also occasionally created simple forms of personal decoration, such as pierced pendants.

The difficult existence of the Neanderthals is reflected in their high frequency of traumatic injury (the remains of all older individuals show signs of serious wounds, sprains, or breaks), abundant evidence of nutritional deprivation during growth (more than 75 percent have evidence of growth defects in their teeth), and low life expectancies (few lived past 40 years, and almost none lived past 50 years). Yet, they were able to keep severely injured individuals alive, in some cases for decades, again reflecting more advanced social organization.

Image of the Neanderthals

The image of the Neanderthals, therefore, is one of archaic humans who shared a number of important characteristics with modern humans, including their large brains, manual dexterity and walking abilities, and social sophistication. Like their archaic predecessors, however, their foraging systems were considerably less efficient at hunting and gathering than those of modern human hunter-gatherers, necessitating larger bodies, more-muscular limbs, greater endurance, and larger faces to hold their large front teeth. It was only with the emergence of modern humans that these archaic features disappeared, being superseded by more elaborate cultural behaviours and technologies.

(E.Tr.)

CRO-MAGNONS

Cro-Magnon is the name of a rock shelter near Les Eyzies-de-Tayac, Dordogne, Fr., where several prehistoric skeletons were found in 1868. Sent to the site, the French geologist Louis Lartet began excavations in which he established the existence of five archaeological layers covered with ash. The age of the human remains found in the topmost layer—along with worked flint and the bones of animals of species now extinct—is Upper Paleolithic (*c.* 35,000–10,000 years ago), but the attribution of these to a clearly defined Upper Paleolithic culture is less definite. Traditionally regarded as Aurignacian, since typically Aurignacian artifacts were found in the rock shelter, they could be more recent, and it has been suggested that they should be assigned to the Perigordian (a separate industry covering approximately the same time period as the Aurignacian), which would give an age of about 25,000 BC.

Examination and analysis of fossils. Although it appears that at the time of discovery the remains of more than 10 individuals existed at Cro-Magnon, only fragments from some five individuals were preserved and studied, among them the cranium and mandible of a male about 50 years old. Considered representative of the Cro-Magnon type, this specimen is known as the "Old Man of Cro-Magnon." Also preserved were skull fragments of about four other individuals, some bits of bone from a fetus or newborn child, and an assortment of bones attributed variously to the individuals mentioned above. The first subject men-

tioned, the Old Man of Cro-Magnon, has been regarded as typical of the Cro-Magnon peoples.

The skull is longheaded (dolichocephalic) and as seen from above has a pentagonal outline, with outward bulging of the parietal bones (at the sides of the skull). The forehead is straight, the browridges only slightly projecting, the cranial vault noticeably flattened, and the occipital bone (at the back of the head) projects backward. The cranial capacity is large, about 1,600 cubic centimetres (about 100 cubic inches). Although the skull is relatively long and narrow, the face appears quite short and wide. This combination is often regarded as a common feature of the Cro-Magnon race. The forward projection of the upper jaw (maxilla) is also distinctive. The eye sockets are low-set, wide, and rather square in shape; and the nasal aperture of the skull is narrow and strongly projecting. The mandible is robust, with massive ascending ramus (the upward projection of the lower jaw, where it attaches to the skull), strongly developed points of muscular attachment, and a quite prominent chin.

Typical skull features

The root of only one molar tooth remains in the jaw of the Old Man of Cro-Magnon, a fact that contributed to the idea of his advanced age. In fact, it is probable that the loss of the majority of his teeth occurred after death. The teeth of the other individuals found at Cro-Magnon, which are similar to the teeth of other fossil humans classed as Cro-Magnon, show that the dentition of Cro-Magnons was nearly identical to that of modern humans. Most of the teeth recovered, however, especially the last molars, are distinctly larger than those of most modern peoples. Dental caries is sometimes apparent, and tooth wear is often extreme.

The remainder of the Cro-Magnon skeleton is not fully known from the remains found at the original site, which are incomplete and poorly preserved. Skeletal material attributed to the Cro-Magnon race from other sites, however, affords the general impression of robustness, probably combined with powerful musculature. The forearm is relatively long, as is the thigh; the femur (thighbone) has a very prominent linea aspera (a bony ridge that runs lengthwise down the back of the femur), and the tibia is flattened from back to front (platycnemy). The hand skeleton is large with short fingers, and the foot has a prominent heel.

Early investigators were impressed by the stature of Cro-Magnon man, as some reconstructions suggest that the Old Man of Cro-Magnon may have been as much as 190 centimetres (six feet three inches) tall. A restudy, however, suggests that the stature of the original Cro-Magnon remains varied from 166 to 171 centimetres (five feet five inches to five feet seven inches). The stature of several skeletons from the Grimaldi Caves (in Italy, near the French frontier), which show clear affinities to those of Cro-Magnon, was noticeably greater, with an average height of 177 centimetres. It is thus reasonable to conclude that, on the whole, the Cro-Magnon peoples were relatively tall.

By courtesy of the American Museum of Natural History, New York

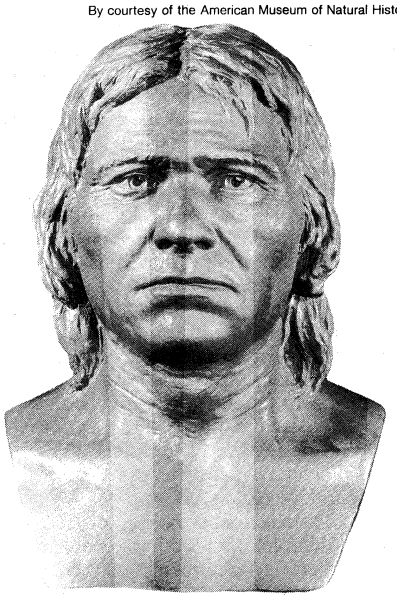


Figure 28: Reconstruction of the appearance of Cro-Magnon man.

Lesions noted in the Cro-Magnon skeletal remains have been attributed to wounds, but one analysis has suggested that these lesions are pathological in origin and may have resulted from the action of a toxic mushroom, *Actinomyces israeli.*

Two French prehistorians, A. de Quatrefages and Ernest Hamy, in 1882 took the Cro-Magnon fossils to be prototypes of a Cro-Magnon race. As opposed to the Neanderthal race—the first remains of which were found about 25 years earlier—the Cro-Magnons were then considered to be the most ancient form of *Homo sapiens.* To the Cro-Magnons were assigned other remains discovered before 1868 at La Madeleine and Bruniquel, in France; Engis, in Belgium; and Paviland, in Wales. Subsequently, further finds of human skeletal remains extended the geographic range of the Cro-Magnon peoples through much of Europe and into Asia and North Africa. Many of the central European fossils, however, belong to a type that differs from the Cro-Magnons called Brno-Předmosti, named for the area of central Czechoslovakia where they were discovered; like the Combe-Capelle remains, which they resemble, these individuals appear to have more primitive characteristics than typical for Cro-Magnons.

Variant Cro-Magnon finds

The place of Cro-Magnons in human evolution. The question of the relation of Cro-Magnons to the earliest forms of *Homo sapiens* is still unclear. It does appear, however, that Cro-Magnons (*H. sapiens sapiens*) and Neanderthals (*H. sapiens neanderthalensis*) are closer in affinity than was once believed. It long was thought that certain Cro-Magnon traits could be seen in human remains of Middle Pleistocene age (900,000–130,000 years old), but this argument no longer seems convincing. The tendency now is to locate the origin of the Cro-Magnon type in western Asia, as typified by the remains found at the Jebel Qafzeh and Skhūl sites in Israel.

Perhaps as complex as the question of origin is that of the duration of Cro-Magnons. It appears that they flourished during the Upper Paleolithic, and that there was a tendency toward more gracile individuals, as seen in the fossils from Saint-Germain-la-Rivière in France. Individuals with at least some Cro-Magnon characteristics—called Cro-Magnoids—are found in the Upper Paleolithic, the Mesolithic (in Europe, *c.* 8000–*c.* 5000 BC)—for example, at Muge, Port.—and in the Neolithic (in Europe, roughly from 5000 to about 2000 BC); at the same time, remains have been found for individuals who were quite different, often brachycephalic (broad-headed). Some modern human groups that are more or less homogeneous are thought to have retained a close relationship to Cro-Magnon types, at least in their cranial morphology. Particularly noteworthy are the Dal people from Dalecarlia (now Dalarna, Swed.) and the Guanches of the Canary Islands, the latter of which is said to represent a relatively pure Cro-Magnon stock.

The culture of the Cro-Magnons. The ties between *Homo sapiens sapiens,* and particularly Cro-Magnon peoples and the various Upper Paleolithic cultures (*e.g.,* Châtelperronian, Aurignacian, and Gravettian, which are classified on the basis of stone and bone tools), are relatively clear, although in 1979 Neanderthal fossils were found in Châtelperronian strata near Saint-Césaire, Fr. It is still difficult to establish precisely an outline of physical types and cultures for this period. Moreover, there are some detectable differences between populations and cultures of western Europe and roughly contemporaneous populations of central or eastern Europe.

Toolmaking. The Cro-Magnon peoples are generally associated with the Aurignacian culture tool industry, and perhaps with the Gravettian (also called Upper Perigordian). The Aurignacian tool industry is characterized by retouched blade tools, end scrapers and "nosed" scrapers, burins (chisellike tools), and fine bone tools, in particular long, flat points (spearheads) with cleft bases. Other bone and reindeer-horn implements are also seen: awls,

Cro-Magnon tool industries

tools for smoothing and scraping leather, and the so-called *bâtons de commandement*—bars of antler or bone with holes drilled in them, the use of which is still uncertain, although they may have been used for straightening arrow or spear shafts. The Gravettian industry differs from the Aurignacian industry in the use of an abrupt retouching technique to form what are called backed blades (*i.e.,* tools with one edge blunted). Modern knowledge of all of these industries has been advanced as the ability to trace how and why various implements were used by Paleolithic peoples has improved.

Dwellings. The dwellings of Cro-Magnons were most often caves and shelters made by rock overhangs, but it is apparent that huts were made also; sometimes these were simply lean-tos against rock walls, but foundation stones and "pavements" of stone in the shape of houses are evidence of complete huts. These houses are not a new development with the Cro-Magnons, however; both the Neanderthals and earlier peoples of the Middle Pleistocene are associated with similar remains. It seems probable that the Cro-Magnons lived fairly settled lives. Studies of occupation sites and the types and extent of remains found in these sites suggest that the rock shelters were inhabited throughout the year rather than seasonally, and it is likely that these Paleolithic hunters moved their homes only when hunting or environmental conditions forced them to do so.

Hunting techniques. The climate in the habitable parts of Europe in Cro-Magnon times was cool to cold. Plants and animals of types associated with tundra and steppe environments were usual. Bone remains found at Cro-Magnon occupation sites indicate that they were successful hunters of such animals as reindeer, bison, wild horse, and even mammoth. As yet, very little is known of Cro-Magnon hunting methods—for example, whether hunting was individual or collective or if bows or traps were used. It is obvious from the animal remains, however, that hunting techniques must have been efficacious.

Aesthetics and religion. Although earlier human groups certainly had religious practices of some sort—the Neanderthal people buried their dead, a practice merely continued and elaborated by Cro-Magnon and later peoples—and no doubt had some appreciation of aesthetics as well, the first examples of prehistoric art are Cro-Magnon. Small engravings, reliefs, and sculptures of animals have been found in Aurignacian and Gravettian sites, as well as a few later statuettes of ivory or stone and occasional engravings in stone of female figures. These figures are usually large-breasted, wide-hipped, and most often apparently pregnant; they are generally assumed to be some sort of fertility symbol, perhaps used in religious or magical rituals intended to promote the fertility of the group or, possibly, of the game.

"Venus" figurines and animal paintings

The Cro-Magnon people also appreciated the decorative aspects of art, as demonstrated by their use of animal pictures and (more often) simple geometric designs to ornament tools and weapons. It is believed that the people of the second half of the Upper Paleolithic—*i.e.,* of the Solutrean and even more so the Magdalenian culture—were of the Cro-Magnoid variety and that they were responsible for the many splendid paintings of animals found in caves in France and Spain; but such sculptures as that called the "Lady," or "Venus," found at Brassempouy, Fr., are thought to be the work of Cro-Magnon artists. (H.J.De.)

HOMO SAPIENS OF ASIA AND AUSTRALASIA

The extinct *Homo sapiens* populations from what was the eastern end of the human range during most of the Pleistocene epoch were once described as the eastern representatives of a "Neanderthal stage" of human evolution. With the realization that the Neanderthals were a unique European and western Asian race, the term archaic *Homo sapiens* has come to describe these earlier members of the species. As a group they are descendants of local *Homo erectus* populations, and many believe that they are the direct ancestors of the more recent and living peoples of East Asia and Australasia.

East Asia. The extinct populations of *H. sapiens* in East Asia evolved from local *H. erectus* populations and share numerous unique regional features with these earliest Asians. The latest remains of *H. erectus,* particularly the Chou-k'ou-tien (Zhoukoudian) H1 mandible and H3 cranium from Layer 3, are dated to 370,000 years ago by amino-acid racemization (a thermoluminescence date of about 300,000 years is reported for the underlying Layer 4). Several of these, particularly the above-mentioned Locus H Chou-k'ou-tien specimens and a vault recovered from the Narmada valley in central India, could conceivably be considered early *H. sapiens,* but the support for this classification is not universal.

Fossil finds in China

The early Asian *H. sapiens* remains are limited to China, where a fairly complete skeleton from Mount Chin-niu (Jinniu) in Liaoning province and a cranial vault from the Ta-li (Dali) site in Shansi are the best-preserved specimens from the late Middle Pleistocene. The Ta-li cranium, which may have been the older of the two, is from a young (less than 30 years old) adult male and combines a thick, massive cranial vault with a robust but short and markedly flattened face. The supraorbitals are vertically thick, curving to follow the orbital contours, and separated from the forehead by a sulcus (furrow). The face below reflects numerous regional distinctions in its transverse flattening, low nasal angle, anterior position of the cheeks, and markedly vertical orientation (involving the entire face from the nasal bones to the premaxilla). Cranial capacity is about 1,120 cubic centimetres (68 cubic inches).

Milford H. Wolpoff

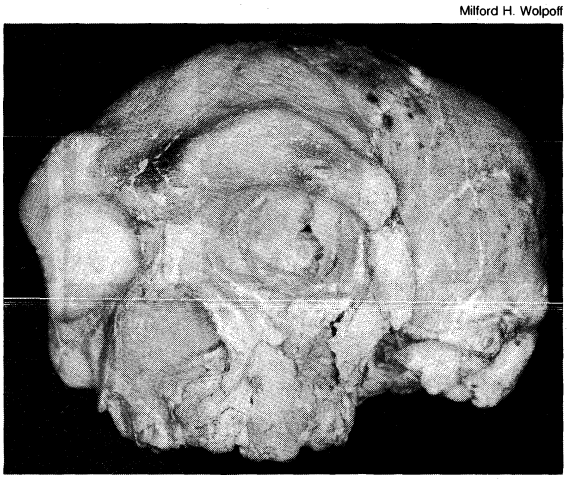


Figure 29: *Homo sapiens* skull found at the Ta-li site in Shansi, China.

Remarkably similar facial features characterize the female cranium (with associated skeleton) from Mount Chin-niu, but the vault is much thinner and about 25 percent larger (1,390 cubic centimetres) and could reflect a younger age. Uranium-thorium dates place the fossils between 200,000 and 100,000 years ago. With the more fragmentary female cranium from the Ma-pa (Maba) site in Kwangtung—which has gnaw marks on the face, suggesting that it was part of an interrupted meal—these two females differ dramatically in robustness from Ta-li while retaining similarities in features common to the region, such as transversely flat and vertically nonprojecting (orthognathic) faces, broad, low noses, arched supraorbitals following the orbital contours, and posterior dental reduction (especially compared with contemporaries to the south; see below). Details of the postcranial skeleton, said to be largely complete, remain unreported.

There are also some more fragmentary dental remains from several sites. The most important features of these remains centre on the unique size, morphology, and distribution of the marginal ridges and basal tubercles on the incisors (*i.e.,* the regionally specific complex of shovel shaping) that characterize these teeth as well as the dentition of many modern Asians.

Earlier Late Pleistocene discoveries of *H. sapiens* are much more fragmentary, and the best-established links for the Middle Pleistocene remains are with the latest Pleistocene finds from the Chou-k'ou-tien Upper Cave, Liuchiang (Liujiang) and Lai-pin (Laibin; Kwangsi), and Tsu-

yang (Ziyang; Szechwan). In addition to the continuation of the Mongoloid features common to the region, discussed above, there are specific resemblances, such as between the brow regions of the Ta-li female and the Chou-k'ou-tien Upper Cave male (specimen 101) and the moderate chignons (occipital buns) of the Chin-niu and Liu-chiang females. Of the earlier remains from the Late Pleistocene, the best-known are the Ch'ang-yang (Changyang; Hupeh) maxilla, teeth from Hsin-tung (Xindong; part of the Chou-k'ou-tien site) and Ting-ts'un (Dingcun; Shansi), and the numerous though fragmentary specimens from Hsü-chia-yao (Xujiayao; also Shansi). Lower facial flattening and orthognathism (Ch'ang-yang and the Hsü-chia-yao juvenile maxillae) and the complex of incisor shoveling are the most important characteristics of the sample.

Australasia. To the south a large sample of early Australasian *H. sapiens* from the Ngandong site in eastern Java (Indonesia)—commonly called Solo man—is clearly linked to much earlier Indonesian *H. erectus* ancestors by morphological comparisons and transitional specimens, such as the cranium from the Sambungmatjan (Sambungmachan) site in central Java. Some workers regard the Ngandong hominids as *H. erectus,* but the features that uniquely distinguish *H. erectus* from *H. sapiens* in Australasia have never been identified; and as the German anatomist and paleoanthropologist Franz Weidenreich (author of the most comprehensive study of these specimens) pointed out, their characteristics mostly fall within the known *H. sapiens* range. That the Ngandong remains are themselves unambiguously ancestral to at least some of the Late Pleistocene Australians is shown by fossil remains from the Willandra Lakes region (New South Wales) such as the specimen designated WLH 50. While their relative temporal sequence is clear, date determinations for all of the Australasian remains are very uncertain.

The Sambungmatjan cranium

The Sambungmatjan cranium, found on the Solo River west of Trinil, is Middle Pleistocene in age. The specimen has many of the special regional features foreshadowing those found in the geologically later sample from Ngandong, a site also on the Solo River but just east of Trinil. The specimen is the faceless vault of an adult male (the sex diagnosis being made on the basis of neck-muscle-related features of the occiput). With a cranial capacity of about 1,100 cubic centimetres, the vault is as large as the largest Chou-k'ou-tien *H. erectus* males and only slightly smaller than the (more recently dated) Ngandong males; it is otherwise similar to them in being long, low, and broadest above the base, with thick cranial bone and a particularly long, flat forehead beginning just behind the straight supraorbitals. Other diagnostic features shared with the Ngandong specimens include a unique thickening at the outer corner of the supraorbitals, which form a rearward-facing triangle, and a flattening at the back of the skull above the prominent nuchal torus.

Later in time, the Middle or early Late Pleistocene remains from Ngandong include two right tibiae and 14 faceless craniums or large cranial fragments that were found close to each other in 1931–33. Two additional fragmentary vaults were discovered nearby on the same high Solo River terrace in 1976, and several others have been recovered subsequently. The six best-preserved vaults (two females, four males) are within the modern range of Australasian size variation and share many regional features with Late Pleistocene and Holocene Australians. Most of the specimens have a supraorbital torus that can be described as a well-developed structure, positioned straight across the orbits except for an indentation at its middle, that is continuous with a long, flattened forehead. There are depressions at varying positions along the middle of the head, generally there is a sharp angulation on the occiput at the nuchal torus—which is prominently developed and separated from the flat surface above it by a pronounced transverse groove—and in many craniums the sphenoparietal articulation is lacking in the region of the temple (on the side of the head). Some of their features seem unusual, or even unique only to this sample. Included in these is the trigone at the corner of the supraorbitals described above and the positioning of the Glaserian fissure just at the roof of the narrow mandibular fossa. The length of the smaller tibia suggests a body height of about 162 centimetres (five feet four inches); the fragment of the larger tibia is from a much taller individual.

Milford H. Wolpoff

Figure 30: The Sambungmatjan cranium, found along the Solo River near Trinil, Java, Indon.

While the Ngandong specimens were not intentionally buried and there are no archaeological associations, some insights into their behaviour can be made. Healed cranial wounds are common and are found on more female craniums than on the males. It has been suggested that the accumulation of faceless skullcaps suggests that they were used as water bowls, and it is uncertain if the limited number of body parts represented is the result of ritual treatment of the dead, taphonomical factors, or the preferences of the collectors.

Links between populations

There are undisputable links between these Indonesian remains and the earliest Australians, populations of modern *Homo sapiens* who are perhaps the first from anywhere to have crossed a water barrier broad enough that one side was not visible from the other. These links can best be seen in the Australian cranium (WLH 50) from the Willandra Lakes region, New South Wales, known to be more than 30,000 years old and suspected of being much older. A direct descendant relationship for this male specimen is indicated by the dimensional and proportional similarities to the Ngandong males and the presence of numerous shared regional features. Many of these features are unique to Australasia and are not found in earlier or contemporary Late Pleistocene humans from other parts of the world. While these regional resemblances have been explained as pathologies, artificial deformations, or "primitive retentions" by some workers, there is no evidence to support any hypothesis of causality except that of ancestry. (M.H.Wo.)

HOMO SAPIENS OF AFRICA

Africa has long occupied centre stage in discussions of the evolution of early hominids. Yet, when considering the origins of the species *Homo sapiens* and in particular modern humans, debate has traditionally centred around Europe. This viewpoint was typified by the American anthropologist Carleton S. Coon, who in the 1960s asserted that Europe and Asia had been the main locations of modern human evolution. In the 1970s and '80s, however, such conceptions were abandoned. Numerous new hominid finds, new absolute datings, and a radical revision of the archaeochronological framework have increasingly placed Africa at the focal point of both *Homo sapiens* evolution and the origin of modern humans.

Fossil evidence. Based upon the fossil finds and dates available for Africa until the late 1960s, it was widely assumed that large areas of the African continent had been populated by very archaic Rhodesoids—named for the cranium found at Broken Hill, Northern Rhodesia (now Kabwe, Zambia)—up to about 30,000 years ago. Many researchers believed that the modern Africans had evolved from these Rhodesoids and had thus developed much later than their modern European counterparts (Cro-Magnons). Some workers even suggested that the modern Africans arose from modern Europeans who had emigrated out of Europe.

Figure 31: Side views of (top) the Kabwe/Broken Hill cranium, representative of early archaic *Homo sapiens;* (centre) the Laetoli LH 18 cranium, representative of late archaic *H. sapiens;* (bottom) the Omo I cranium, representative of early anatomically modern *H. sapiens.*

© Gunter Brauer

Toward the end of the 1960s several important finds were recovered: the remains of three individuals from the Kibish Formation of the Omo River in Ethiopia and the remains of a number of individuals at the Klasies River Mouth site, on the southern coast of South Africa. These specimens, most of which were anatomically modern, were determined to be some 130,000 to 80,000 years old, a finding that astonished the workers and did not fit well with the ideas of the period.

It was only after additional fossils were found during the 1970s, however, that conceptions about the evolution of *H. sapiens* in sub-Saharan Africa actually began to change. This shift was aided by the fact that the dates of many long-known finds were revised and determined to be significantly older than previously had been assumed. Based upon a new analysis of the total fossil evidence, the West German anthropologist Günter Bräuer was able to demonstrate a continuous evolutionary sequence in sub-Saharan Africa during the last 400,000 years. Three divisions have been differentiated in this sequence: early archaic *H. sapiens* (*c.* 400,000–200,000 years ago), late archaic *H. sapiens* (*c.* 200,000–100,000 years ago), and anatomically modern *H. sapiens* (*c.* 100,000 years ago–present). The course of *H. sapiens* evolution in North Africa is not as well known; for this reason, the fossil evidence from this area will be discussed after that from the sub-Sahara.

Three divisions of *H. sapiens* evolution

Sub-Saharan Africa. Early archaic *H. sapiens* from sub-Saharan Africa is represented particularly by the hominid finds at Bodo (Ethiopia), Ndutu and Eyasi (Tanzania), Kabwe/Broken Hill (Zambia), and Hopefield/Saldanha (South Africa). This division possesses some primitive *H. erectus*-like features along with derived features of *H. sapiens.* Cranial capacity is generally greater than 1,250 cubic centimetres (76 cubic inches), and the lateral walls of the cranial vault are more or less vertically oriented, similar to the modern type. The primary primitive features consist of a strongly developed supraorbital torus, an angular occipital, and a typically quite massive face. Even these features, however, also display clear trends in the modern direction. In spite of the considerable variability among these finds (thought, in part, to be the result of sexual dimorphism) and their relatively small number, it may be assumed that this division of *H. sapiens* was widely spread throughout Africa.

The late archaic *H. sapiens* of sub-Saharan Africa consists especially of the Omo II (Ethiopia), Ileret (Koobi Fora) KNM-ER 3884 and Eliye Springs ES 11693 (Kenya), Laetoli LH 18 (Tanzania), and Florisbad (South Africa) hominids. While these specimens display individually varying mosaics of primitive and derived features, they nevertheless possess a form that is distinctly more modern than that of early archaic *H. sapiens.* This is reflected in the increased cranial capacity (greater than 1,350 cubic centimetres) and the usually modern form of the face. Even where archaic traits are present, *e.g.,* in the supraorbital tori, these are considerably reduced in comparison to the early archaic condition. Obviously, such hominids as LH 18 or Florisbad stand very close to the threshold of anatomically modern humans.

The evolutionary transition to modern humans may be seen in both the morphology of the late archaic *H. sapiens* specimens as well as in certain archaic reminiscences among some of the earliest modern finds. Among the most important members of this division are the hominids Omo I and the remains from the Klasies River Mouth and Border Cave (South Africa) sites. The Omo I cranium, which is remarkably similar to the Cro-Magnoid craniums from North Africa and Europe, may certainly be classified as anatomically modern. In spite of their considerable variability, the cranial and mandibular remains from Klasies River Mouth are modern as well, although some of these fragments also appear to exhibit influences from late archaic populations. The cranium and the mandible from Border Cave (specimens 1 and 2) are also fundamentally anatomically modern. It must be noted, however, that these early moderns generally do not display any clear relationships to current human populations, and thus most likely represent a less differentiated, more pristine form of modern *H. sapiens.*

Transition to modern humans

North Africa. The fossil record in North Africa (nearly all of which has been found in Morocco) is less clear, and it is therefore difficult to connect the evolutionary sequence for this area with that of sub-Saharan Africa. There are several reasons for this: only a few finds, most of them jaw fragments, have been obtained from the Middle Pleistocene (Sīdī ʿAbd ar-Raḥmān [Sidi Abderrahman], Thomas Quarries, Rabat); there are uncertainties in the evaluation of the possibly pathological Salé hominid; the dates for the Jebel Irhoud hominids are uncertain; and the influences that the Eurasian Neanderthals may have had upon the Late Pleistocene North Africans are still the subject of debate. If the Jebel Irhoud hominids are indeed

found to be of early Late Pleistocene age (see below), then it is conceivable from a morphological point of view that there were indeed close connections to the sub-Saharan late archaic *Homo sapiens,* although this does not rule out the possibility that there may also have been some influence from the Neanderthals of the Mediterranean area. The most likely date for the appearance of early modern humans in North Africa is between 70,000 and 40,000 years ago, as represented by the finds at the Dar-es-Soltane II cave site (Morocco).

Revision of fossil dates

Dating the fossils. In the early 1970s a number of new carbon-14 datings gave rise to a radical revision in the then-accepted dating of the African Stone Age. The resulting picture has since been confirmed by a variety of newer methods. Accordingly, the African Middle Stone Age no longer is considered to have begun some 40,000 years ago and ended *c.* 10,000 years ago but instead to have begun *c.* 200,000 years ago and ended some 40,000 years ago. One result of this enormous expansion was that much greater ages began to be accepted for a number of hominids, including those found at Kabwe/Broken Hill, Hopefield/Saldanha, and the Cave of Hearths (South Africa). Beyond this, a number of important reexcavations helped to further clarify many dates that had long been uncertain.

After F.H. Smith and F. Spencer (eds.), *The Origins of Modern Humans: A World Survey of the Fossil Evidence,* copyright © 1984 Alan R. Liss, Inc.

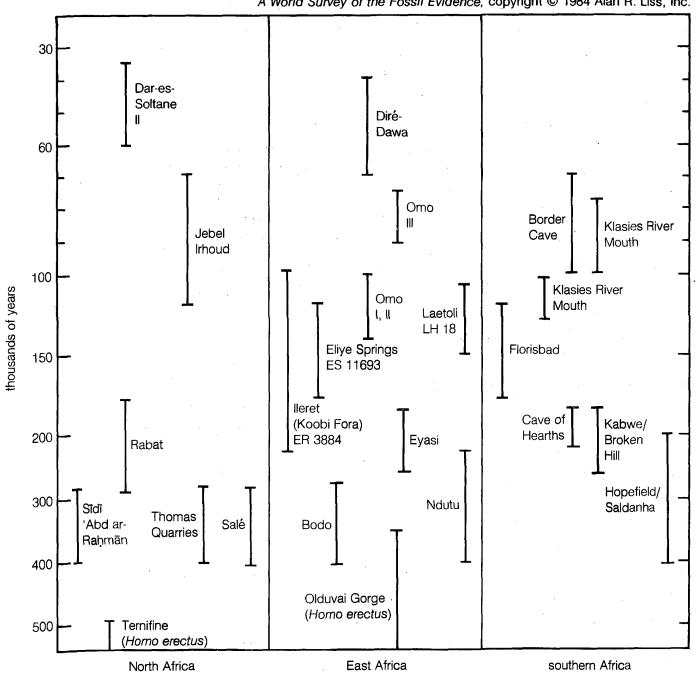


Figure 32: The age spans and chronological positions of the major African *Homo sapiens* fossil finds.

Referring to the representatives of early archaic *Homo sapiens,* the British prehistorian J. Desmond Clark has suggested that all of these finds most likely date from between 400,000 and *c.* 200,000 years ago. Reexcavations and new absolute dates have also yielded support for the ages now accepted for the late archaic specimens. A reinvestigation at Florisbad demonstrated that the hominid found there is between 200,000 and 100,000 years old. For LH 18 the uranium–thorium method has yielded dates ranging from about 130,000 to about 100,000 years. In addition, the stratigraphic provenance of the KNM-ER 3884 hominid indicates that this specimen is probably more than 100,000 years old.

Subsequent research also has provided confirmation for the greater ages of the early modern finds. Reexcavations and various redatings at Klasies River Mouth yielded an age from somewhat more than 100,000 to 80,000 years ago for the modern human remains associated with the African Middle Stone Age. On the basis of all of the available evidence, the German-born geographer and anthropologist Karl Butzer considered the *c.*-130,000-year age for the Omo I find (determined using the uranium–thorium method) to be quite likely. Because of the matrix present on the Border Cave 1 hominid, Butzer suggested that this specimen is associated with a level in the deposits that is about 90,000 years old.

While the dates for important finds representing all three divisions of sub-Saharan *Homo sapiens* are quite reliable, those for a number of North African finds are less certain. In contrast to earlier assumptions, which usually considered the Jebel Irhoud hominids to be between 60,000 and 40,000 years old, electron spin resonance dates now indicate that these specimens are at least 75,000 years old. Dates for the North African Aterian industry also vary widely, ranging from more than 70,000 to less than 40,000 years ago. The modern hominid from Dar-es-Soltane II, associated with the Aterian industry, may thus be between 70,000 and 50,000 years old.

Techno-logical changes

Behavioral inferences. The revision of the archaeochronological framework for sub-Saharan Africa during the 1970s demonstrated that great technological changes occurred in parallel both north and south of the Sahara and thus that the Middle Stone Age of Africa was broadly contemporaneous with the Middle Paleolithic of Eurasia. A comparison of the technologies of sub-Saharan Africa with the fossil record of the region indicates that early archaic *H. sapiens* were still generally associated with late Acheulean assemblages (*e.g.,* at Hopefield/Saldanha and Eyasi), while late archaic *H. sapiens* were associated with the Middle Stone Age (*e.g.,* at Laetoli). On the other hand, the transition to the early modern humans was not characterized by any recognizable technological innovation but took place during the Middle Stone Age. Precisely how the modern human form arose some 100,000 years ago and in what regard the behaviour of these early moderns may have differed from their predecessors is difficult to explain because of a lack of corresponding archaeological evidence. In all likelihood, changes were involved that would not have been preserved in either the fossil or the archaeological record. At least some of these changes may be attributable to the full development of human language and the associated increase in intellectual abilities. In addition, improved social systems and demographic changes probably also contributed to the success of the modern humans.

Evolutionary implications. Conceptions of the course of evolution of *Homo sapiens* have changed drastically since the 1960s. No longer is Africa considered an also-ran in the global context; instead, it has become the region in which the evolution of *H. sapiens* up to modern humans is best documented and where, moreover, modern humans appeared earlier than in any other part of the world. The implications that this changed view has for the origins of modern humans outside of Africa remain the subject of much controversy. Some researchers, such as the American anthropologist Milford H. Wolpoff, have interpreted the African lineage as a regional phenomenon, arguing that other lines of development were responsible for the rise of the modern humans elsewhere. As far as Europe and East Asia are concerned, however, others have expressed strong doubts about the evidence for such regional transitions to modern humans. The present discontinuities between archaic and modern humans in the fossil record outside of Africa and the later appearance of modern humans in those areas have instead led to the view that modern humans originated solely in Africa; subsequently, from there they radiated to the Middle East and then to Europe, supplanting the Neanderthals living in these regions. It can be assumed, however, that during this process some intermingling between these populations also took place.

Although the fossil record is less clear for East Asia, there are indications that the moderns from Africa also expanded into that region. Molecular biological studies by the New Zealand-born American Allan C. Wilson and his associates support the view that present-day human populations had their common origin in Africa, even though the extent of gene flow that may have occurred between the regional archaic populations and the radiating modern populations during the global expansion of modern humans remains a point of much debate. (G.Br.)

MODERN HUMAN POPULATIONS

General considerations

All living human populations belong to a single biological species (*Homo sapiens*) within a larger group or genus (*Homo*). Within the human species a large number of populations may be differentiated genetically through readily observable characteristics (*e.g.*, skin, hair, and face and body proportions) and through less obvious but more distinctive biological traits, such as blood type. These biological groupings within species are commonly called races, in humans as well as in other living forms.

RACE AND POPULATION

Definitions and terminology. The term race as applied to humans has been variously used—by politicians, military leaders, philologists, human biologists, demographers, and historians. Some "races" constitute language groups, often of peoples whose only kinship is that they speak a common language. Such was the original meaning of the so-called Aryan race. Some "races" are simply hypothetical, invented to embrace present distributions of such genetic (hereditary) characteristics as stature or hair colour—*e.g.*, the Nordics. (The word Nordic also has been given a political meaning, referring, despite their differences in physical characteristics, to peoples in northern Europe.) Race has been variously applied to national or cultural groupings, as in the days when English writers referred to an Irish race and to a Scottish race. As used in census and other applications, the designation race often groups different peoples for administrative convenience; thus, the category Hispanic may group people from Meso-America, the Caribbean, South America, and the Philippines who may differ considerably in their racial origins.

"Race" also has been applied to human groups inferred to have existed on the basis of archaeological discoveries; the Etruscan race is an example. Various religious groups who may or may not have common ancestry sometimes are called races—the Jewish race, for example. By extension of biblical thinking and in honour of Shem, son of Noah, a Semitic race was conceived in an effort to describe peoples who spoke Semitic tongues, some of whom may have learned their language more recently than others.

All of those uses of the term race are separate and distinct from its biological meaning in classification (taxonomy)—the natural divisions or groupings below the species level. As such, race differs from breed or line, which refer to artificially established groups maintained by intensive selection or by deliberate hybridization. Just as the term race is often too broadly applied to the entire species of man (as in the human race), particular race names invented to explain distributions of observable physical characteristics of human populations are not biologically meaningful.

The misuse of the word race—particularly the manner in which it was employed by Nazi Germany—had led workers to search for alternate terms. Some biological descriptions refer to human stocks, one intention for this being to avoid political overtones. Other writers have favoured the word division in lieu of race, again apparently to escape what may be perceived as offensive connotations. Other references to these human groupings include strain (without implying the equivalent of purebred strains of laboratory animals); variety (although the specific botanical meaning does not apply to human races as ordinarily constituted); and ethnic group, which, although generally meaning cultural or political groupings (*e.g.*, Macedonians, Croats, Magyars, or Slovenes), is at times used with exactly the same biological meaning as race. With the advent of population genetics, establishing gene frequencies in specific populations, many workers have come to prefer the word population for taxonomic purposes. Populations so defined, however—such as San (Bushman), Ainu, Lapp, Eskimo, Coloured (South Africa), or Micronesian—are often the same groupings that have been or can be called races. Still, population is a useful addition for such linguistically and genetically distinct groups as the Basques and is an easier concept to explain.

Other terminology

The term geographic, or continental, race is often used to describe populations that occupy a broad geographic range. Likewise, local race is used for populations in a more restricted area, and microrace may correspond to a single, extended breeding population. These natural groupings, which reflect geographic (and therefore reproductive) isolation, display a range of genetic differences that are the focus of much research. The ultimate questions are how long the races (or populations) have been distinguishable and what processes brought about the distinctions.

What the different geographic races are called is to some extent unimportant as long as the same terminology is employed by all; such traditional designations as white, yellow, and black, however, are clearly inappropriate. The designations for local races and microraces are similarly unimportant, except for communication and for the sensitivities of the people themselves. It has long been a practice on the part of some human taxonomists to convert placenames into taxonomic names by adding the suffix *-id* (*e.g.*, Pennsylvanid, Montanid) or the suffix *-oid* (Capoid for the Cape peoples of South Africa). Geographic terms, without suffixes, also suffice (hence, the Mediterranean race) or are used in conjunction with language groupings, where justified (*e.g.*, Azteco-Tanoan). Often reference is made to

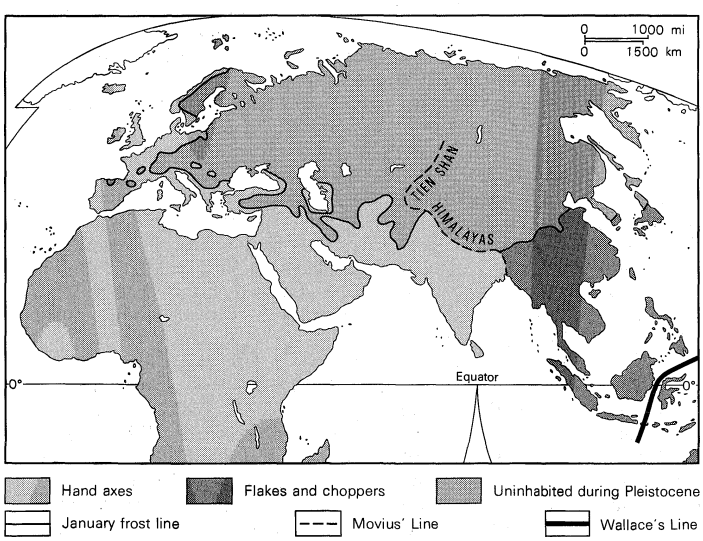


Figure 33: Distribution of human groups through habitable regions of the world in the Late Pleistocene.

particular national or cultural groupings, such as Finns, because available data are so arranged, or to artificial groupings (*e.g.,* Ghanaians or Vietnamese) until further information can establish more precisely the makeup of those groups. Even a designation as being from a city may not be enough, given demographic and genetic differentiation within cities, Tokyo being a typical example.

Geographic races. Naturally occurring (*i.e.,* produced by natural, usually geographic separation of human groups) races of the human species are by no means identical in number of members or degree of genetic differentiation. There are small groupings of a few hundred to a few thousand individuals, some slightly and only recently isolated reproductively from adjacent people. Other equally small groups may have been mating apart from the rest of mankind for centuries or even for thousands of years. Members of some human races number in the hundreds of millions (as the peoples of modern Europe) or in the billions (as in Asia).

It is both useful and meaningful to identify the very large human groupings that often correspond to continents or other major geographic areas as geographic races, a term extensively used with other life forms. Geographic races are numerically large, containing within them smaller groups of reproductive isolates (breeding populations).

Reasons for a geographic delineation

The reasons for the large groups' geographic delineation are usually clear. The Indians of the Americas were reproductively separated from the peoples of other geographic regions for many thousands of years. Thus, they have come to differ genetically from the rest of mankind and even from those Asians from whom they stemmed. The Australian Aborigines similarly constitute a geographically defined group of local races (see below) separated for millennia from the rest of the world, except for some slight contact, until the late 18th century.

Local races. Most geographic races include numerous local breeding populations. While such local races may have few members, others number in the hundreds of millions. Local populations (local races, usually with distinctive languages) dotted pre-Columbian America; there were hundreds in North America alone. In other parts of the world, where high population densities have built up, local races are also apparent, as in many parts of Africa. History provides a picture of ancient local races in much of the European continent, with numerous local groupings in the British Isles cataloged and categorized by Julius Caesar. In the course of the last two millennia, however, local races in Europe seem to have become less easily defined, except along such broad lines as northwestern Europe, the Alps, and the eastern Mediterranean. Such local races are the ones most easily identified and studied; they constitute population units that display continuing evolutionary change. Many of these units are allopatric in that they live in distinct places and their boundaries are easily defined. Others are sympatric—that is, they live in the same municipality or region but preserve their genetic identity. Hopi living in enclaves on Navajo lands or in government-assigned territories and their Navajo neighbours each constitute nearly allopatric populations, whereas enclaves of Korean nationals in Los Angeles or Algerians in Paris constitute distinct sympatric populations.

Microraces and smaller units. For much of the world, local races exist as reproductively isolated, culturally distinct, and genetically differentiated breeding populations. Indeed, barriers to gene flow (interbreeding) are prerequisites for the continuing existence of these populations. Culture and geography together tend to maintain such hereditary differences over time; thus, social or caste prohibitions on interbreeding serve to maintain local races in India. Elsewhere, religious regulations have maintained many local races for millennia. Rivers were efficient barriers to interbreeding in older days, and oceans remain major barriers to intermarriage. As populations have grown, however, as in Europe, geographic distance itself has become less of a barrier, and distance alone does not have the isolating influence it once had.

Yet sheer density of population is itself a restriction on gene flow and on the distance over which human mating is likely to occur. A man is less likely to venture far afield for marriage when there are many single women within the same apartment block in which he lives. Under such circumstances, the human male has the same effective mating range as the mosquito. Thus, even in heavily settled locations, genetic differentiation continues in its course. In a long-established city such as Tokyo genetic differences occur from district to district; this also holds for Rome and London and other large cities that have been studied. Some of these differences clearly represent older residence patterns of breeding that support the continuance of localized groups of people (microraces) long genetically distinct (as in the Limehouse district of London or in certain of the ghetto districts of Rome). In moving from one coast of England to the other, there are found systematic genetic differences, some related to earlier settlements and cultural patterns, still others apparently reflecting later differential directions of selection.

Causes of gene-flow restrictions

Associations with race. *Adaptations to environment.* Local genetic adaptations may well be expected for races long occupying the same habitat, accustomed to a distinctive way of life, and used to a particular climate. Long before the microscope made it possible to count melanin granules in skin sections, naturalists had come to appreciate the importance of differences in skin (and eye) pigmentation. They observed a direct relationship between sunlight levels and skin pigmentation, at least in Europe and Africa. Later it was determined that the pigment granules serve as a natural sunscreen, protecting the deeper layers of skin as harmful ultraviolet wavelengths in sunlight are absorbed, and that there is a correlation between higher melanin content and slightly lower sweating thresholds. Truly confirming experiments would be both unethical and impractical, but it has been demonstrated that the rate of skin cancer rises sharply as pigment densities go down; in Texas, for example, skin cancer is far more common among light-skinned and freckled individuals. Light-skinned people may generate considerable melanin with continued exposure to sunlight (and so tan), but this does not occur until after damage to the skin has been done. Those who have less melanin in their irises are also more likely to develop cataracts from long-term exposure to ultraviolet radiation. Conversely, low levels of melanin in the skin appear to be an advantage in parts of the world where cloudy conditions are common or where (as in the polar regions) there are prolonged periods of limited sunlight. Lighter-skinned individuals are more able to convert vitamin D precursors in the skin to usable vitamin D. This theoretical advantage has been confirmed: rickets and osteomalacia (progressive mineral loss from bones), both disorders caused by vitamin D deficiency, are more common among dark-skinned children living in northern Europe, with its generally cloudier climate and shortened hours of sunlight in winter.

Differences in body build or bodily proportions also interested early naturalists, and subsequently it has been shown that these differences are also adaptive at thermal extremes. At higher air temperatures a lanky build allows for greater heat loss—by convection, conduction, and radiation—and thus a large surface area relative to body mass makes for greater comfort. When air temperatures are low the trend is toward a small surface area relative to mass, which tends to conserve body heat. These adaptive traits have been demonstrated in environmental laboratories by measuring the threshold of shivering as temperatures are lowered and the increase of body temperature when the environment is heated. It is therefore of interest in population comparisons that the surface-to-mass ratio is lowest among some Asian and North American Eskimos and highest among such African groups as the Nilotes. The relative size of the extremities (hands and feet) also enters into consideration: short hands and pudgy fingers are less susceptible to frostbite than are long hands with thin fingers, especially if there is also increased vascular flow.

Foods

Potentially intolerable or toxic substances in foods constitute another area of influence by the environment. Lactose, the complex sugar in milk, is poorly absorbed by the adults from most populations because as they get older they lose the ability to produce sufficient lactase, the enzyme that breaks lactose down into simpler sugars.

Northern Europeans generally maintain high lactase levels into adulthood, but even among Europeans some infants do not have high enough lactase levels and are therefore intolerant even of their mothers' milk. This is a highly inadaptive circumstance, and such infants cannot thrive unless the genetic defect is recognized and appropriate substitutes are provided. Some infants lack the ability to metabolize other sugars, which is also a serious detriment. At the other end of the spectrum, wheat gluten and other glutens are poorly tolerated by some, particularly in western Europe (and especially Wales and western Ireland). Wheat has been a staple for thousands of years in Anatolia, the Middle East, and North Africa; it is a "late" food in northern Europe. The Windsor (fava, or broad) bean is a common staple food in southern Europe, but some individuals, particularly those of Mediterranean origin, develop hemolytic anemia (excessive destruction of red blood cells) if they eat them. The Egyptians were aware of this disorder, as were the Romans, and in modern times it has been identified as an inherited trait.

Many foods contain substances that are toxic—such as cyanide compounds—or substances that can render particular nutrients unusable, although the level of toxicity or unusability varies widely. The phytates in wheat can render zinc and iron biologically unavailable; this can result in growth failure and dwarfism, as has been observed in the Middle East. Oxylates in spinach and other greens can inhibit the absorption of calcium, and although the Cucurbitaceae (which includes cucumbers) are generally good sources of several vitamins and may help prevent cancer, some individuals are intolerant of them. The ability to metabolize ethanol also varies among populations and individuals, and these differences may be adaptive or inadaptive, depending upon social circumstances.

Diseases

Population differences in disease susceptibility and resistance to infectious diseases were long postulated, largely on anecdotal evidence. Africans were imported into the plantations of the New World, for example, on the assumption that they were more resistant to malaria. It is recognized that there was some truth in this belief, now that the sickle-cell gene and the red-cell enzyme "defects" have been identified. It is also known how such genes spread in Africa, the Middle East, some parts of Indonesia, and even into southern Europe (Italy). A wide variety of diseases, some of which have been around for a long time, are now being considered for their population implications. For example, the incidence of tuberculosis, which peaked in the mid-19th century, has declined since through improved sanitation and possibly because of improved nutrition; but it is also possible that the virulence of the responsible microorganism (the tubercle bacillus) diminished, the pool of genetically susceptible individuals was reduced, or these two factors worked in combination. Syphilis is another disease that seems to have changed in both virulence and symptomology over time. Similarly, there has been much speculation about the role of these factors in the outbreak and subsidence of the world's great epidemics. More recently, attention has focused on the human immunodeficiency virus (HIV-1), the causative agent of acquired immunodeficiency syndrome (AIDS). It has been suggested that the precurser to the HIV-1 virus was present in a relatively benign state in Africa for a long time and that it mutated into the more virulent form; also under exploration has been the possibility that individuals as well as populations may exhibit differing degrees of susceptibility to the virus.

Language and population. It was once conjectured that languages were originally influenced by hereditary configurations of the mouth, teeth, and tongue of their earliest speakers. Research, however, has failed to relate the characteristics of any spoken language to the genetically transmitted facial configurations of its speakers. English, for instance, seems to be no better learned by children of English ancestry than by those of German, Italian, Egyptian, Ethiopian, or Pakistani derivation. There is an increasingly popular theory, however, that all languages, like all races, have a common origin and that some words are understandable to all peoples, regardless of the language they speak. This hypothesis has been tested by asking children to match words and meanings in a cross-cultural and cross-language context. The results of these tests have been arguably positive, suggesting that vocabulary may have a neurological basis.

Correlation between language families and geographic races

Although language is learned and not genetically inherited, there is an obvious correspondence between the major language families and the geographic races, both of which are the result of initial differences, long maintained, intensified, and directed by thousands of years of geographic separation. Thus, linguists look to genetic differences between groups as an adjunct to linguistic studies, and human geneticists take an interest in comparative linguistics. The smaller or lesser divisions in each geographic area tend to speak related languages, as can be discerned in Denmark, Germany, The Netherlands, part of Belgium, and England. Language as well as gene frequencies identify the Navajo as being of Canadian origin, distinguish the Basques from the French-speaking and Spanish-speaking peoples surrounding them, and separate the Finns from their other Scandinavian neighbours. Conversely, there may be much diversity within a language grouping. Though written Arabic has spread over much of Africa and the Middle East, spoken Arabic still reflects older language and genetic distinctions. Likewise, the dialects of Chinese differ markedly between regions while the written language is mutually intelligible. English in Northumberland and across the Scottish border strongly reflects the language of the Gaelic speakers and follows differences in blood-group distributions.

In Europe there is a strong relationship between genetic distance and linguistic distance. Such groups as the Finns and the Basques are most distant from others, while Celtic speakers (in Ireland, Scotland, and Brittany) are less distant from English and Frisian speakers in their genetic makeup than in their languages. In addition, one language may supplant another as the result of political subjugation. This can be traced from both genetic and linguistic evidence, the latter in the form of loanwords and expressions. English, for example, is full of loanwords of French and Danish origin. Family names (surnames) also indicate previous periods of genetic and linguistic interchange as, for example, the Danish-derived names Anderson and Johnson in England; it is more difficult to trace how the French name Beauchamp became Beecham or how the Spanish name Henriques became Hendricks, but the historical and genetic evidence is there.

The languages and peoples of the Pacific Basin also provide collaborative evidence of population movements. Linguistically and genetically, the Polynesians preserve evidence of their origins in and relationships to the Indonesian archipelago. Some words, particularly those relating to specific cultivars, have spread so far and so rapidly from their places of origin that it is very difficult to trace their origins.

Race and "intelligence." People of one group traditionally have found reasons to disparage people of other groups on the basis of their behaviour, language, and other cultural attributes. Thus, the Arabs of the Middle Ages were highly critical of the Frankish (French) traders who did not bathe regularly, as were the Chinese of English mariners. The English took a dim view of many foods in India (although they did borrow chutney), and the Japanese were reluctant to accept beef until the keeping of cattle and the eating of beef was made acceptable by official edict. Commonly, people of one group have regarded people of another group as less trustworthy and, especially, less intelligent. Victors often described the vanquished as being of lower intellectual capacity, and colonizers considered the peoples they came to live among as inherently less capable, mentally inferior, and more childlike. There were learned theological arguments in the past as to whether the natives of North America even had souls, and equally learned discussions (in Europe) on whether or not these people had the capacity to read, write, and do sums.

Such notions were comforting to subjugators in the era of colonialism, and they were particularly congenial when slavery became economically important in the Americas. Yet subjugated peoples often had technological superiority over their new rulers, as shown by the metalworking,

weaving, and survival skills of Africans that were highly appropriate to their native habitats. The English-speaking settlers of North America were highly ethnocentric, according the highest intellectual capacities first to themselves, then to Scottish immigrants, and then (to a lesser extent) to those of German and Scandinavian origin who followed. Later immigrants from Ireland, eastern Europe and the Balkans, and Italy were accorded lesser abilities and capacities. Intelligence tests that were first put into use in the first two decades of the 20th century seemed to confirm earlier beliefs: people who spoke English as a native language scored highest, those who spoke other Germanic languages came next, and those recently arrived from eastern and southern Europe attained much lower test scores. It was later recognized that scores on these tests were language-dependent and also measured familiarity with the culture. For example, the descendants of people from Asia (who scored low on the tests in the early 20th century) have become some of the highest scorers. Another factor now realized is that long exposure to a culture of poverty depresses school-oriented test scores; Americans who have been impoverished for generations continue to score low on the tests.

Inaccuracy of intelligence tests

One may well ask whether people who have long lived in inhospitable deserts, in Arctic and subarctic climes, or in rain forests could have been of inferior mental capacity. Such conditions place a premium on survival skills, demanding a deep knowledge of botany, animal behaviour, and climatology.

GENETIC FACTORS AFFECTING HUMAN POPULATIONS

Admixture. Throughout human history, genetically different human populations have met and mixed. In many instances admixture (also called miscegenation, or gene flow) was accomplished through the friendly meeting of peoples, while in other instances it came as a consequence of invasion and conquest. Often treaties were sealed by an exchange of marriageable members, creating instant kinfolk on both sides of a political boundary. While warfare, invasion, and military occupation might suggest that gene flow was usually unidirectional—from the victor to the vanquished—reverse gene flow also took place. The Vikings contributed their genes to Ireland, England, and especially Scotland, but they also took captive brides home, altering their own gene pools in the process. American occupation forces in Japan after World War II fathered thousands of Japanese-American children, but they also took thousands of Japanese wives back to the United States; subsequent military excursions into Korea and Vietnam have also added genes in both directions.

England, of course, has long experienced gene flow, from the Romans and from Ireland (especially in the north) and Denmark. Spain experienced the advent of Roman legions and then occupation by the Arabic-speaking Moors. Sicily still shows the results of Arab occupation, and the Balkans still contain Muslims, the consequence of long Turkish domination.

The North American melting pot

The North American melting pot is perhaps the prime example of admixture potential, since so many of the world's genetic populations have been brought together in the United States; very few Americans can claim ancestry from fewer than three nations. There are, however, places in the world, such as Brazil and Hawaii, where admixture in relatively recent times is even more complicated. Quite often the ancestors of individuals from these areas came from two or even three continents. A typical Hawaiian, for example, may be an assemblage of Portuguese, Hawaiian, Japanese, and Irish genes. These examples, drawn from older history and more recent international excursions, mark instances of admixture for which there are good records. Earlier historical accounts, which may be less detailed, still reflect the opportunities for admixture over the centuries, such as in Europe from the Hun, Mongol, and other "barbarian" invasions. Few people can claim truly "pure" descent, even over a few centuries and even when travel was at a rate of only a few miles per day. Modern air transport now makes gene flow possible to any point on the globe in a matter of hours.

Isolation. Isolation, either geographic (in the form of oceans, mountain ranges, and deserts) or cultural (factors that favour gene exchange within groups) preserves racial differences and allows for genetic adaptation to climatic and disease factors over long periods of time. Geographic barriers are not perfect, of course, as individuals do cross them, and endogamy is rarely complete. Still, these barriers have remained effective in preserving particular sets of genes within populations. Pre-Columbian America is the classic case of genetic isolation by geography. Except for a small amount of gene flow across the Bering Strait, the peoples of the Americas were almost completely isolated from the genetic changes in Europe, Africa, and Asia and were thus maximally responsive to selective pressures in the New World; the severity of this responsiveness was best demonstrated in the indigenous Americans' lack of immunity to European diseases. Europe, on the other hand, is more of a case of cultural isolation. Geography was never as confining to the south or east. Nevertheless, vast expanses of desert and water and distance served as major impediments to gene flow when the rate of travel was only a few miles per day. As populations increased in size and territory, the possibilities of gene flow also increased, especially as previously separate groups newly came to have contiguous frontiers. But increased population densities and a nearly continuous distribution of humanity from the Baltic to the Adriatic actually tended to impede the flow of advantageous genes, since individuals did not have to travel far to find a mate.

Mutation. New genes in a given population arise by mutation and at a rate proportional to the level of penetrating radiation. Thus, the mutation rate is probably higher in Denver, Colo., than it is in Dubuque, Iowa (in this instance because Denver receives more solar radiation than Dubuque), in people who have been subject to medical X rays, and in those exposed to fallout from atomic weapons. Whether or not a mutant gene becomes established and spreads in a population depends on several factors. If it is both dominant and deleterious it is soon lost, though it may be first expressed as a spontaneous abortion or a developmental defect. If the mutant gene is recessive it can increase in frequency in the population, and if it is advantageous it will increase in frequency under the influence of natural selection. It has been possible to trace some mutant genes, the classic case being the gene for hemophilia that Queen Victoria passed through her descendants to other royal houses of Europe.

Tracing mutant genes

Mutant genes are the raw material upon which natural selection acts. Many of the differences between populations and races presumably arose as mutations were eliminated by selection in some cases and retained in other cases because of their degree of advantage at some place or period in time. For most of the morphological differences that can be seen and measured the advantages are often difficult to adduce, but differences in enzyme or hemoglobin levels are more accessible to investigation. It is clear that the mutant genes for the various atypical hemoglobins (*e.g.,* hemoglobin S and hemoglobin E) were immediately advantageous, since they conferred immunity to malaria, although they were disadvantageous where that disease was absent. The history of mutations culminating in lesser skin pigmentation, however, can only be guessed; it is clearly disadvantageous in areas of high solar radiation but more nearly neutral or possibly advantageous in cloudier climates.

Small-sample effects and human diversity. When populations are small and the number of breeding pairs smaller still, there can be large fluctuations in gene frequencies from generation to generation. Called small-sample effects, they are of particular interest because they provide one explanation for genetic differences among populations. Some genes may be lost from a small gene pool, and other genes may become fixed at a frequency of 1.0 simply by chance. This is the small-sample effect called genetic drift, which is of particular importance when the population size is 1,000 or less and the number of breeding pairs as small as 100. Most early preagricultural and incipient agricultural populations were of necessity that small or smaller; many modern hunter-gatherer groups are still no larger. It is likely that these early populations exhibited

wide fluctuations in gene frequencies from generation to generation, which may account for some of the geographic and local differences that are seen in modern populations.

Moreover, as populations enlarge and split into new populations, purely random allocation of genes rarely occurs. Historically, those who moved away and formed new groups tended to be related and so might differ genetically from those left behind. It is known from the genealogical records kept by the Polynesians that their migrant groups consisted of close relatives; and, when Amish groups in the United States split, the separate populations differed from the start. These are examples of what is called the founder effect. The effect taken to its extreme is evident on Pitcairn Island, where most of the living Pitcairners are descendants of Fletcher Christian, the leader of the HMS *Bounty* mutineers. Similarly, genealogical records kept by the Mormon church document the reproductive success of Joseph Smith and a few others. Family names and historical records also demonstrate how much reproductive success on the part of some of the founders can affect the genetic makeup of small towns and even some larger cities. Though family names usually are patronymic and therefore ignore the women in successive generations, they can be useful measures of microevolutionary change within defined populations in relatively recent times.

The founder effect

THE STUDY OF HUMAN POPULATIONS

Human populations are distinguished or classified in terms of genetically transmitted differences. They are studied (in terms of their hereditary origins and their biological relationships with other such races) for evidence of ongoing evolution and of continuing genetic change. Such studies help explain the origin and persistence of genetically determined diseases and serve to explore their long-term influence. For races that have been long resident in their present locations, research sheds light on the long-term genetic effects of such environmental factors as temperature and climate and of food type, source, and availability. For races that have been formed in the recent and historical past, the investigations can reveal the proportion of different groups that entered into genetic admixture. For particular groups with common racial ancestry that have moved from their original locations, evidence of genetic change also interests researchers.

For the most part, investigations of populations involve groups of individuals rather than individuals themselves. Differences and similarities between groups are expressed as trait frequencies, gene frequencies, and frequencies of different amino-acid sequences; from these frequencies, admixture—both ancient and recent—can be estimated.

Early methods of study. *Morphological comparisons.* For centuries, geographic and local races were identified by the most obvious physical differences, chief among them the colour of eyes, skin, and hair. From such observations came the simple notion of a few groupings based on the apparent colour of the skin alone. In the 19th century, attempts were made to create racial classifications based on the form of the hair: straight, curly, woolly. To some extent these classifications have validity; but hair form is affected by modes of hairdressing, and there is great individual variability in unmodified hair form within groups in many parts of the world. Other observations included those for eyelid form, nasal shape, and cranial shape. The assumption in all these cases was that agreement in trait frequencies was an indication of taxonomic affinity.

Anthropometry. The development of the technique of anthropometry (human bodily measurement) in the 19th century brought other approaches to the identification of races. Various groups of "pygmies" were identified by their extremely short stature. The proportions of leg and torso (trunk) were a distinguishing mark of short-legged groups such as the Eskimo and southern Asians and long-legged peoples such as the Nilotes.

Head measurements (craniometry) showed distinctively roundheaded (brachycephalic) populations like those of central Europe and longheaded (dolichocephalic) populations such as those typical of northern Europe. Cranial and facial bone measurements helped to distinguish differences in living people and in skeletons alike and clarified such features as the broad faces and jutting jaw angles of northern Asians and the broad-faced, narrower heads of some American Indian groups.

Body measurements do reflect a genetic component, but they only can be used with caution, since they are also affected by nutrition. Better-nourished people tend to grow larger in many dimensions because they have more fat and because a more ample diet enables them to build more muscle and bone. In regions in which the nutrition of the population is relatively uniform, however, anthropometry can be useful in comparing genetically isolated subgroups. Anthropometric comparisons are also useful in the study of the skeletal remains of ancient populations, and they are irreplaceable in archaeology for the study of fossil remains.

Modern methods of study. Since about 1940, workers in human taxonomy have studied factors that are genetically and simply determined, such as single-gene traits. While genetically complex differences as in the size and shape of the teeth can also be measured, their utility is limited by ignorance of their exact mode of inheritance. Thus, wisdom teeth never develop (third-molar agenesis) among an exceptionally high proportion of Eskimo and in many Asians, while, by contrast, third-molar agenesis is exceptionally rare in Australia, New Guinea, and in some parts of Africa. This genetically complex trait, however, is relatively little used in making racial comparisons. There are simply inherited skeletal traits, such as presence of a broad middle segment of the fifth, or little, finger (brachymesophalangia-5), that are common in some populations and extremely rare in others. Fusion of two of the wristbones occurs in up to 6 percent of individuals in some parts of Africa but in only the smallest fraction of the populations of Europe or Asia. Such skeletal differences can be studied readily with X rays but are not so easily detected in field studies through ordinary visual inspection. Significant racial differences in the size and thickness of the walls of longer bones also require X rays for their measurement.

Little-used particular traits

Single-gene traits. The greatest amount of information on simply inherited traits bearing on race has come from the study of the blood and from biochemical analysis of the urine. While the components of the blood are no more informative than are those of muscle or viscera, many millions of people throughout the world have had their blood typed, producing an accumulation of information on blood differences, their inheritance, and their geographic distribution. In similar fashion, modern medical interest in metabolic abnormalities ("biochemical lesions"), often diagnosed by measuring amounts of different amino acids in the urine, has revealed various biochemical polymorphisms (a variety of forms or types) that characterize families of close relatives, microraces, local races, and, in some cases, geographic races as well.

Some findings about race have been by-products of other kinds of study. It was discovered almost by accident that a chemical called phenylthiocarbamide tastes bitter to some people but seems tasteless to others. Only later did physical anthropologists and human geneticists discover that some racial groups are remarkably "taste-blind" to phenylthiocarbamide. Population differences in colour blindness were discovered in the course of routine, large-scale visual examinations for military purposes. Earwax polymorphism (dry-flaky and moist-sticky) was long known in Japan; this observation led European and African geneticists, who had assumed all earwax was sticky, to discover the polymorphism in their own lands. Discovery of the genetically determined deficiency of the enzyme glucose-6-phosphate dehydrogenase (G6PD) came about when new antimalarial drugs were introduced after World War II and occasionally caused dangerous side effects in persons of African and Middle Eastern origin.

Blood traits. The classic human blood groups are examples of traits produced in the individual by the action of a single set of genes within the set of chromosomes (the majority of genetically determined features of the individual come about through the interaction of many gene sets). The first blood groups to be elucidated were those of the ABO system. Because it is known how the A, B, and

The role of blood groups

O groups are inherited, the frequencies of their genes in any human population can be calculated from the results of blood tests. For blood group B, for example, the gene frequency is close to zero among American Indians and Australian Aborigines and occurs in as many as 40 percent of the people in parts of Africa and Asia; it is only a minor blood group in Europe, the observed frequency rarely exceeding 12 percent. Apparently, Australia became isolated reproductively from the rest of the world before B became frequent elsewhere. Most investigators assume either that the East Asian ancestors of American Indians crossed into the North American Arctic before blood group B had become established in Asia or that, having had it once, they lost blood group B in the process of evolution.

Other blood groups often used in racial comparisons include the MN series. The frequencies of M and N are about equal in most parts of Europe, and they do little to characterize local differences there. But since M is virtually missing in Aboriginal Australia and rare in the other lands of the Pacific Basin, and since N is absent or nearly so in American Indians, M and N have great value in comparing Pacific populations, in postulating their origins, and in assessing European or African admixture in various American groups.

Within the Diego blood group, the Diego positive gene (Di^a) is especially common among people of Asian origin. It also has been detected among people of Polish origin in the United States, perhaps as a residue of the Mongol invasions of Europe. Yet Di^a cannot be used to compare people within groups that lack the gene, nor is the question of origins easily resolved simply by comparing those who have it in the highest frequency. Not every genetically determined set of traits has equal value in comparing, differentiating, or searching for the origins of each race. Great differences within the same geographic race point to local selective factors and considerable blood-group diversity. Even within small and related villages, investigators do not use the same list of genetically determined traits for every set of people studied.

Other blood factors

Blood factors other than blood groups include abnormal red pigments, such as hemoglobin S (the variant hemoglobin that is the cause of sickle-cell anemia), which is rare in most parts of the world but found in as many as half the local residents in some parts of Africa. Inherited blood factors include blood substances called haptoglobins and transferrins and deficiency of the red-cell enzyme G6PD, mentioned above. Dozens of additional properties of red blood cells and of other blood fractions have become known through widespread medical use of blood transfusions. Among the best-known of these properties are the so-called Rh blood subtypes, first discovered in rhesus monkeys and later in humans. These became of medical importance in Europe when incompatibility between the Rh types of the mother and her unborn baby was found to damage or kill the fetus. The troublesome Rh-negative gene (r) is relatively common in Europe (especially among the Basques) but less frequent elsewhere; thus, its frequency can be used as one measure of European admixture in populations that once lacked it. Rare or absent in the rest of the world, R_0 (another distinctive gene in the Rh series) is relatively common in Africa. Thus, the occurrence of R_0 can provide a measure of the gene flow into the mixed population of Brazil or out from the "black" people of the port of Puerto Barrios in Guatemala. By its relative frequency, R_0 serves as one measure of the admixture of European and African genes in the United States.

Other genetic traits. Besides blood fractions and blood groups, there are genetically determined differences in the excretion of amino acids via the urine. While some people excrete much β-amino-isobutyric acid (BAIB), for example, the urine of most people shows very little of this substance even when all share the same diet. Genetically determined BAIB polymorphism and any metabolic advantages it may provide are incompletely understood. Nevertheless, people who do show high levels of BAIB excretion are Asians and some American Indians. If the trait is assumed to have come from Asia, it is an unexpected finding that many Indian groups living nearest to Asia (theoretically of more recent arrival) seem to excrete less BAIB than those living in South America.

Congenital defects

Through continuing natural selection, the frequency of congenital defects differs among populations. Phenylketonuria, a cause of mental deficiency that arises from an inherited enzyme disorder, is far less common in Africa and Asia, for instance, than it is in Europe. Jews from near Vilnius, Lithuanian S.S.R., tend to a relatively high frequency of Tay–Sachs disease (progressive blindness and mental defect in infants); people from England are particularly prone to develop a bone disorder called Paget's disease; and congenital hip dislocations are especially common in American Indians from the Southwest.

Certain types of cleft palate are remarkably common among Japanese, while abnormal narrowing of the body's largest artery, the aorta, is far less frequent in Americans of African origin than in their fellow citizens of other origins. Minor wristbone (carpal) fusions are much more common in people of African ancestry than in Mexican-Americans, who in turn show broad middle segments of the fifth finger much more frequently than black Americans. Women of African origin are comparatively unlikely to suffer the spontaneous fractures and low-back pains that come from adult bone loss (osteoporosis); yet men of African origin who are military pilots are unusually prone to suffer ejection-seat fractures of the lower lumbar vertebrae. Painfully impacted wisdom teeth are most common among Asians and Europeans who are genetically delayed in third-molar development.

Much of the increasing ability to calculate genetic distances and genetic similarities and to infer ancestral relationships is a serendipitous product of advances in medical knowledge. For example, the necessity of matching donors and recipients in organ and tissue transplants resulted in the discovery of the genetically determined human leukocyte group A (HLA), the main transplant antigens. Genetic mapping, involving amino-acid sequences, holds the greatest promise to human evolutionary research because so many different genes can be sampled. A product of genetic engineering, it can help explain how different human populations have been separated and how different lines have converged or combined.

The races of mankind

THE ANTIQUITY OF RACES

Homo sapiens is currently viewed by most systematists to be a single polytypic species derived directly from its predecessor species, *H. erectus,* through continual transformation directed by natural selection. As such, *H. sapiens* has a single origin, in terms of its single predecessor species, but with geographic differentiation arising well before the start of the Holocene epoch (about 10,000 years ago). At least some of the geographically delimited divisions of the species (*i.e.,* geographic, or continental, races) owe many of their distinctive features to differences present at an earlier time that were preserved. For other geographic races, intermediate status in morphological and skeletal features, consistent with their historic locations, may be due to admixture during the Pleistocene epoch or to competing directions of selection.

Modern dating methods support the belief that at least some of the geographic races have been in situ since at least the Late Pleistocene and possibly earlier than that. People resembling living Europeans appear to have lived in North Africa, in the Mediterranean, and even northward before the Holocene. Sub-Saharan Africa appears to have been inhabited by peoples resembling its current inhabitants since the Late Pleistocene. Australoids, or people much like them, may have been in residence in Australia for as long as 30,000 years. The evidence is convincing that the ancestors of the American Indians arrived via land bridges from Asia before the end of the last glaciation. Thus, it is probable that some of the geographic races actually antedate the emergence of pre-sapient hominids. The notion that some of the skeletal (and biochemical) differences between geographic races may antedate the *H. erectus–H. sapiens* boundary is both exciting and debatable. It contradicts the older assumption that *H. sapiens*

Possibility that races antedate *H. sapiens*

arose first, then migrated to different parts of the world, and then differentiated into the modern groupings.

The smaller groupings (local races and microraces) are for the most part of post-Pleistocene origin. Throughout the world, however, there are isolated populations of unique character. Unless these groups possessed extraordinary means of sea travel in prehistory, it must be assumed that they also originally migrated via the system of land bridges that connected continents during the last glaciation.

Human populations, such as American blacks, are of such recent formation that historical records and serological data can be used to trace their origins and to calculate the rate and degree of admixture with Europeans. Similar information has also been applied to the study of the racially complex populations of Brazil, many of the islands of the Caribbean, and Hawaii. In parts of the world where there are less complete historical records of migrations and invasions, serological data can be used to estimate, for example, the extent of Mongol influence in eastern Europe, Turkish influence in Greece and Yugoslavia, and Moorish influence in Spain and Sicily.

It is possible to justify a small number of geographic races if the number of defined populations is suitably restricted. Inclusion of more discrete populations raises the number of geographic races somewhat. It is also possible to identify both a very large number of local races, identical to the number of discrete populations, or to a smaller number placed in historical perspective. Not every population has been explored, and it is possible (but not probable) that additional geographic races will be described in the future. While new additions to technology, such as genetic mapping, will improve, alter, and amend the knowledge of human taxonomy and of relationships (*i.e.*, phylogeny), such additions are likely to be incremental.

THE GEOGRAPHIC RACES

The notion that there were but three primary races of mankind (Caucasoid, Negroid, and Mongoloid) became untenable during the European age of exploration, when new continents, new island chains, and new landmasses were discovered (along with new populations that did not fit into one of the three groupings mentioned above). American Indians, though resembling their Asian neighbours in some respects, were distinctly different in other respects. The Aborigines of Australia, called blacks by Europeans because of their dark skins, differed markedly from the peoples of equatorial Africa. So did the dark-skinned Melanesians, who were so named because of their skin pigmentation. The Polynesians, who resembled mainland Asians in some respects, were not exactly Mongoloid. Likewise, the peoples of the Indian subcontinent, though resembling their neighbours in Persia to the northwest, were not quite Caucasoid.

There are now thought to be between six and 10 major geographically delimited groupings, depending upon the assumptions made in defining them and the degree of attention given to their numbers. It is legitimate, for example, to separate American Indians and Asians because of the millennia that separate them in time, but it is also legitimate to group them (despite differences in blood-group frequencies and the absence of blood group B in American Indians). The same dichotomy exists between Australian Aborigines and Melanesians. Nonetheless, there is a broad consensus among workers for designating nine historical geographic races. They are (1) the African geographic race, which consists of sub-Saharan Africa; (2) the European geographic race, which includes Europe, North Africa, and the Middle East; (3) the Asiatic geographic race, which includes Central, East, and Southeast Asia and the Aleutian Islands and western Alaska; (4) the American Indian geographic race, which includes all of North and South America except western Alaska and the Aleutian Islands; (5) the Indian geographic race, which includes the Indian subcontinent to Nepal and the Iranian border; (6) the Australian geographic race, including Australia and (formerly) Tasmania; (7) the Polynesian geographic race, including the island arch defined by Easter Island, the Hawaiian Islands, and New Zealand; (8) the Micronesian geographic race, including the islands of Yap, Pohnpei, and Guam, and with continuities with the Polynesian group; and (9) the Melanesian geographic race, which includes the island of New Guinea with continuities with the Australian group (see Figure 34).

Number of geographic races

Even a listing of nine geographic races leaves a number of isolated local populations in taxonomic limbo. In Africa these include the Mbuti (Pygmies) of the Ituri Forest of Zaire and the San (Bushmen) of the Kalahari. In Asia the Ainu of northern Japan and groups in Taiwan and mainland China do not fit neatly into the category Asiatics. Eskimo in Asia and North America complicate any enumeration, for, although they are conveniently included under Asiatics, some are actually American Indians who have adapted an Eskimo way of life. Similarly, the Negritos of the Philippines are not easily included in either the Asiatic or Micronesian groups. In addition, there are numerous groupings of historically recent and hybrid origin throughout the Caribbean, in much of Central and South America, and in North America that complicate any attempt at a tidy taxonomy.

Of the nine geographic races, four have received more study than the others, both in morphological and genetic terms. These four are the African, European, American Indian, and Polynesian groupings. They are described in more detail below.

The African geographic race. Africa is geographically isolated from the rest of the world by oceans and seas around its perimeter. Only the narrow and inhospitable Sinai Desert connects the continent to the Middle East. Further isolation is provided by the Sahara, which separates much of Africa from its Mediterranean coast. Thus,

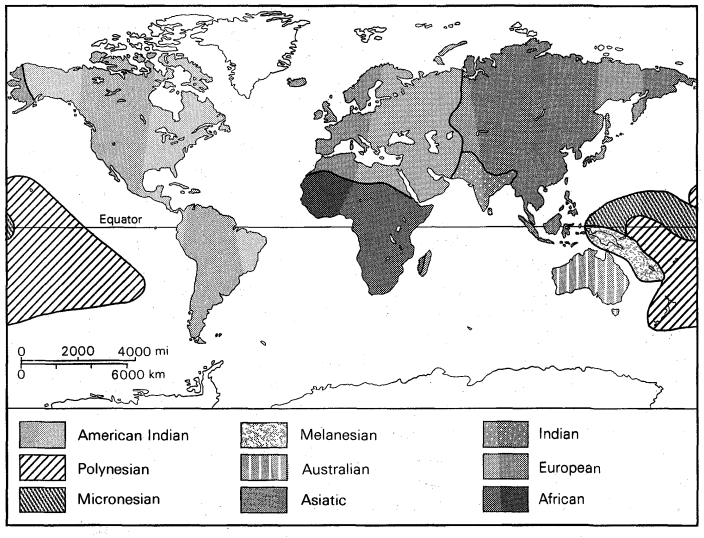


Figure 34: Historical locations of the major races of man.

most of the peoples of Africa constitute a single geographic race but with much local diversity, some of it extending to the last glaciation and before.

Africans south of the Sahara are characterized by heavy concentrations of melanin in the skin. Hair form in much of Africa is distinctive, growing in spiral tufts and with limitations on the attainable length. Tooth crown sizes tend to be large and third-molar agenesis rare. Both tooth formation and tooth eruption tend to be more rapid than in Europeans. Serologically, sub-Saharan Africans are characterized by a rather high incidence of blood group B, by some geographically limited variants in the Rh series, by the S blood group, especially around the Cape of Good Hope, and in the more malarial areas by hemoglobin S and a different constellation of red-cell enzymes. Most African peoples are low in adult lactase levels, except for those long habituated to the use of cow's milk. Geographic diversity is shown in various ways, including lighter pigmentation in the Cape area in the far south and in Ethiopia and Somalia nearest the Gulf of Aden. The present distribution of peoples in Africa has been much influenced by the introduction of such crops as corn (maize) and the potato, which has permitted more intensive agriculture and—by reducing the forest cover—more open areas and clearings, which have favoured the spread of malaria-carrying mosquitoes. Hemoglobin S, advantageous in the presence of malaria, has probably spread in relatively recent times.

Homeland of *Homo*

Africa, especially the area around Lake Tanganyika, now appears to be the homeland of *Homo,* in which case all humans must be of African ancestry. The greatest quantities of prehominid and australopithecine fossils have been found in Africa, and it is possible that the physical appearance and skeletal and serological makeup of present-day Africans is of very considerable antiquity. Thus, it raises the question of just what features were typical of the species before the migrations to Europe and Asia, and which features have come about—in Africa—after humans spread further to all parts of the globe. The Mbuti of the Ituri Forest are of particular interest as an ancient size adaptation with a distinctive variant of growth hormone. The San of the Kalahari are also a distinctive group, remnants of a much larger population, whose pictographs are found over a considerable area of the continent.

The movement of peoples out of Africa in historical times generally has been slow. In ancient times the Nile River served as a conduit for Africans from the Sudan to the cities of Egypt. The greatest outflow occurred between the 16th and 19th centuries, when several million equatorial Africans were captured and taken as slaves to the Americas. Millions were also taken, in servitude, to the Arabian Peninsula.

The European geographic race. Before Iceland, the Americas, Australia, and other places were colonized by Europeans, the European geographic race corresponded in its range to the European continent, North Africa, and the Middle East, with its outermost limit in Ireland to the west and Turkistan to the east. Except for Ireland, a relatively late settlement, these same borders were occupied by distinctively European-appearing peoples for millennia, extending back to the Upper Paleolithic.

Though differing in systematic fashion from north to south and from the northwest (Northern Ireland, the Scottish Highlands, and Denmark) to the southeast (including the Arabian Peninsula), the living peoples of this group are characterized by a rather high frequency of the Rh-negative blood type c, a low frequency of blood group B, nearly equal frequencies of groups M and N, and a variable incidence of relatively unpigmented hair and reduced skin pigmentation (which increases on a northwest–southeast axis). The ability to continue lactase production into adulthood is quite "European" but again declines on a northwest–southeast axis. In general, people from Europe, North Africa, and the Middle East tend to be later in tooth formation and eruption than are other groups.

There are enclaves in Europe that fall out of this description, notably the Lapps of Scandinavia and the Gypsies (who originated in India), and there are groups with linguistic origins outside of Europe proper (*e.g.,* the Finns). But despite population movements and migrations, both overland and by sea, people of European ancestry in many areas have retained skeletal similarities with their Neolithic and even Upper Paleolithic ancestors.

The American Indian geographic race. In facial form, tooth morphology, and minor peculiarities such as brachymesophalangia-5, the indigenous peoples of the Americas often resemble Asiatics, indicating a common ancestry. Some, in fact, have maintained affinities with the Asian mainland through the present, as is true for the Aleuts of the Aleutian Islands and the Eskimo of northern Alaska. Yet most of the "Indians" that Christopher Columbus met had been residents of the Americas for thousands of years, dating back to the existence of land bridges and a more salubrious climate in Alaska and Siberia. It is therefore appropriate to note the Asian origins and similarities, and the differences that have come about with time and isolation by distance and altered sea levels.

Asian origins

The absence of blood group B in American Indians is notable, suggesting that they came to America before blood group B had attained a high frequency in Asia. In the MNS series, American Indians are overwhelmingly M, again indicative of the long separation. Yet such genetically determined traits as shovel-shaped incisors and a relatively high frequency of third-molar agenesis again reaffirm Asian origins, as does eyelid morphology and (for most Indians) projecting malars (cheekbones). Hair form tends to be straight, hair shafts tend to be rather thick, and body hair and facial hair tend to be sparse. Bearded representations are virtually absent in graphic art from the northwest coast, Central America, and Peru. American Indians appear to share with Asiatics a low tolerance for alcohol, indicated by lower levels of the alcohol-degrading enzyme alcohol dehydrogenase.

Most American Indians do not look like "schoolbook" (Plains) Indians, and many do not look especially Asiatic, for variability is great over the range of two continents. Though some groups developed high civilizations and then disappeared, there also has been considerable continuity in many areas with living populations closely resembling those recovered from ancient burials. Many American Indian groups were eliminated as a result of disease, colonial expansion, and forced relocation; but others have come to exceed their pre-Columbian numbers, with varying degrees of admixture from Europe and Africa.

The Polynesian geographic race. The Polynesians were still expanding their territory when the British explorer James Cook first encountered them, and their overseas migrations extending from near New Zealand to Easter Island can be traced. Much of their social structure and their ability to navigate over thousand of miles of open waters is known, and museums preserve many of the resplendent feather cloaks and ceremonial objects they made. European mariners marveled at the Polynesians' body size and appearance. Though they had some Asiatic traits and a distinctively Asiatic appearance, they also appeared quite European.

It is possible to think of the Polynesians in terms of multiple origins—*i.e.,* Asiatic × Melanesian × "primitive" white. It is equally possible to think of the Polynesians as a purely distinctive group of their own. Over thousands of years (as suggested by ancient burials), many now-unoccupied Pacific islands were populated by Polynesians; this fact is reflected by the spread of their cultivars and by the depletion of indigenous animal and bird populations. It is a moot point whether the Polynesians ever reached coastal South America, though it is possible. It is certain, however, that their population numbers were once far larger, and that their language helps to identify their earliest origins.

LOCAL RACES AND MICRORACES

Besides the relatively few geographically delimited groupings just mentioned, there is also a very large number of genetically distinct populations that occupy a defined territory and often speak a language of their own. Many of these local races have been isolated for thousands of years and owe their unique genetic composition to reproductive isolation, selective forces peculiar to their climatic location, their mating practices, and the chance combination of genes that marked them at the start. The Ainu

of northern Japan are of particular interest because they represent the early preagricultural inhabitants who preceded the Neolithic invaders. Other groups—including the San of the Kalahari, the now-extinct Tasmanians, and the people of Tierra del Fuego at the southern tip of South America—are of interest, both genetically and culturally, because they exemplify the population dynamics of numerically small populations with limited mate selection, shortened life spans, and, often, changing gene frequencies from generation to generation.

Local races in larger populations

Local races are not limited to the most distant places and the most hostile environments (*e.g.,* Belgium has both Flemish-speaking and French-speaking populations), but distance and isolation have preserved such groupings over long periods of time. The Basques are of interest because of their unique language (unrelated to Indo-European languages) and their distinctive blood-group gene frequencies. The Berbers in North Africa, who speak a distinct language and who long antedate the spread of Arabic-speaking peoples, were in situ long before the spread of agriculturalists in that part of the world.

The great number of local races and their diversity complicates any attempt at simple listing, such as the numerous "Eskimo" groups, some of which are derived from American Indians and others of which are related culturally, linguistically, and genetically to the Eskimo now spread throughout the Arctic. American Indian groupings, many of which are valid local races, present a wide variety of language groupings and distinct gene frequencies, and it is often possible to re-create their migrations over time. It is known, for example, that the Navajo derive from northern Canada and that they are only relatively recent arrivals in the American Southwest, where they now encircle Hopi groupings that are the remnants of pueblo dwellers who long resided in that area.

In Europe with its hundreds of millions and Asia with its billions, it is often difficult to establish genetic boundaries for smaller populations. Marked average differences in many hereditary traits over the regions from northern to southern Europe, however, have been plotted like lines on a weather map; these clearly show enclaves of genetically distinct peoples. The Basques, and the Lapps of Scandinavia, represent such old enclaves. Gypsies, who still speak a language of Indian origin, are spread across Europe. In addition, European Jews preserve their Middle Eastern origins despite intermarriage with the local populations and their historical movements. Thus, the Sephardic Jews who left Spain during the Inquisition and the Ashkenazi Jews, especially those who live in enclaves in eastern Europe, preserve distinctions; and the Cochin Jews of India are distinct from these two groups and more closely resemble Jews from the Arabian Peninsula, who, in turn, share far more genes with Ethiopian Jews.

In Asia there are many geographically defined local races. Koreans remain distinct, as do the Japanese—although they are somewhat less uniform, especially if Okinawans are included. Differences between the Tibetans and the Han (Chinese) are considerable, and in China there are many local groupings that long have been distinct.

Local races often developed where agriculture, especially the cultivation of high-yielding cereal and tuber crops, was introduced at a late date. Groups that were wholly or largely dependent on hunting and the availability of game and that often lived in adverse climates could not maintain large populations. Even after the introduction of agriculture, yields often remained scanty: Scottish Highlanders considered themselves lucky to harvest four grains of rye for each seed planted, and before the introduction of the potato, agriculture did not support a large population in Ireland. That so many locally defined and genetically distinct populations have continued to exist in such a densely populated area as Europe can be understood in terms of the food supply. Although Britain profited much from industrialization after 1800, for example, its population expansion was fueled by relatively cheap imported grains and ameliorated by emigration to North America and Australia. Wick and Tyrie remain serologically distinct from East Anglia and Eastbourne, as they have for thousands of years.

THE EVER-CHANGING RACES

Population size and ongoing evolution. Population size is a critical factor in ongoing human evolution, ultimately determining whether any given population will continue to exist. If the number of breeding pairs of a population is too small and if many of them are infertile, it will disappear as a separate entity, even if it continues by merging with another. If, on the other hand, the number of breeding pairs becomes large and successful in reproduction, the number of mouths may exceed available resources. In the latter event the population must split and spread, forming at least two daughter populations and extending its territory as well.

Availability of mates

When populations are small, there may be little choice of mates and scant opportunity for selective mating. There may be large disparities in the ages of spouses, as is the case for some small Eskimo groups. Among the living Samaritans of Palestine, historical records reveal that preferential cousin marriage was not practiced when mate availability was low, but that it returned as the population increased. In addition, the smaller the size of a population, the less likely it is that any given mutation will appear, assuming a constant mutation rate over time. A given mutation is also more likely to be lost in a small population simply by accident, although at the other end of the probability curve a given mutation may also achieve a high gene frequency in a small population and even become fixed. These attributes are of interest to medical geneticists in particular, who search for clues to juvenile diabetes, hypercholesterolemia, and other chronic, genetically determined diseases.

Large populations with a thousand breeding pairs (and numbering in the tens of thousands) are also of interest because purely chance effects are less likely to affect gene frequencies from one generation to the next, thus allowing the classic mechanisms of natural selection to function unhindered. Also, large populations are likely to include a considerable socioeconomic range, so that the genetic effects of being poorer or being richer can be measured. In theory, a large population has built-in protection against the accidents of nature and of humans. The Aleuts of the Aleutian Islands were decimated after the arrival of the Russians in the late 18th century, but hundreds of them still survive (their numbers even increased somewhat after Alaska became part of the United States), while smaller groups on the Alaskan mainland now have few survivors. An originally large population size has helped to preserve the Armenians after their original lands were dispersed and despite the losses suffered in Turkey. Dispersion has also served to protect numerous other groups, including the Lapps and European Jews. In each such situation, localized groups have come to show genetic differences due to accidents of sampling and equally localized selective factors.

Large populations tend to acquire smaller, marginal groups, which may lose their identities unless they become localized within limits set by the larger group. Ultimately, the result is an increase in the gene pool, as is now seen in the United States. Often only names remain to identify the original contributors, as with the Scottish, German, Irish, and Polish surnames in a typical Midwestern industrial city—and many of these are Anglicizations and convenient misspellings. Numerically large populations, however, are not necessarily uniform genetic units; identifiable subgroups often suffer some form of discrimination, although it is often the case that those who are discriminated against may enjoy a reproductive advantage over the discriminators. Large class differences in fertility can therefore exist even in a "classless" society, which can give rise to important microevolutionary changes over time.

The ingredients of population evolution

Thus, ongoing evolution at the population level is a product of chance, geography, and natural selection. It also is affected by religion and local attitudes toward reproduction—*e.g.,* whether such attitudes support a norm of large or small families. Between populations, the rate of human evolution may be a question of technological advancement, of the availability of food, or of major medical advances that reduce the number of deaths. As countries develop improved food technology, more effective food

distribution, immunization, and other medical care, their populations tend to increase, thereby changing the genetic makeup of the world's population.

The future of races. The human species is in the process of accelerated evolutionary change, brought about by alterations in the relative size of different populations and by the breaching of geographic and social barriers to gene flow. When hundreds of thousands of individuals are transported from South Korea and Vietnam to the United States, from Pakistan and India to England, or (as guest workers) from Turkey to West Germany and Denmark, gene frequencies in the host countries can scarcely stay static. As China and most of Europe approach zero population growth while Central and South American numbers rapidly expand, it is evident that the human gene pool will be vastly different in a hundred years.

Changes in climate have affected population growth in the relatively recent past. The Little Ice Age in Europe, a cold period that ended about a century ago, had an adverse effect on agriculture. In the 20th century drought in eastern Africa and flooding in Bangladesh have had serious consequences for populations in those regions. The postulated greenhouse effect, which may increase global temperatures and lead to altered patterns of drought and rainfall, holds implications for the future of food production and the relative numbers of people in different geographic populations.

Until the 20th century, malaria was a potent selective force in much of the world, sparing those who lived at higher latitudes and altitudes and those who possessed atypical hemoglobins and red-cell enzyme defects; it continues to be a selective factor in some areas. The HIV-1 virus has become a major selective force in some parts of Africa. New diseases arise, often as mutant forms of less virulent maladies, and with increased international contact can rapidly sweep across the globe. At the same time, some old diseases, such as tuberculosis, seem to have lost their virulence, and thus their role as a possible selective factor is reduced.

Despite these actual or potential events, it is likely that the geographically defined human groupings that emerged from the Pleistocene will remain distinguishable for a long time to come, since these groups still have the advantages of relative geographic separation and vast numbers. It is also likely that many of the locally distinct populations will persist into the foreseeable future, although some do teeter on the brink of disappearance. Political events will determine, for example, whether the Kurds—who have been known since biblical times—survive and whether the San will be assimilated into the Bantu speakers or continue as a distinct group. (S.M.G.)

BIBLIOGRAPHY

General: CHARLES DARWIN, *The Descent of Man and Selection in Relation to Sex,* 2 vol. (1871), is historically the foundation reference. EDWARD O. WILSON, *Sociobiology: The New Synthesis* (1975); G.E. KENNEDY, *Paleoanthropology* (1980); and MILFORD H. WOLPOFF, *Paleoanthropology* (1980), are comprehensive general books.

Overviews of human evolution include JOHN BUETTNER-JANUSCH, *Origins of Man: Physical Anthropology* (1966); J.R. NAPIER, *The Roots of Mankind* (1970); W.E. Le GROS CLARK, *The Fossil Evidence for Human Evolution,* 3rd ed. rev. and enlarged by BERNARD G. CAMPBELL (1978); BERNARD WOOD, *Human Evolution* (1978); RICHARD E. LEAKEY, *The Making of Mankind* (1981); C.B. STRINGER (ed.), *Aspects of Human Evolution* (1981); BERNARD G. CAMPBELL, *Human Evolution: An Introduction to Man's Adaptations,* 3rd ed. (1985); ROGER LEWIN, *In the Age of Mankind: A Smithsonian Book of Human Evolution* (1988), and *Human Evolution: An Illustrated Introduction,* 2nd ed. (1989); D.R. PILBEAM, "Human Evolution," Part I of G.A. HARRISON *et al., Human Biology: An Introduction to Human Evolution, Variation, Growth, and Adaptability,* 3rd ed. (1988), pp. 1–143; JOHN READER, *Missing Links: The Hunt for Earliest Man,* 2nd ed. (1988); PAUL MELLARS, "Major Issues in the Emergence of Modern Humans," *Current Anthropology,* 30(3):349–385 (June 1989); and FRANK E. POIRIER, *Understanding Human Evolution,* 2nd ed. (1990).

IAN TATTERSALL, ERIC DELSON, and JOHN VAN COUVERING (eds.), *Encyclopedia of Human Evolution and Prehistory* (1988), is detailed and comprehensive and includes up-to-date bibliographies. A detailed, authoritative, and highly useful catalog of fossil finds is MICHAEL H. DAY, *Guide to Fossil Man,* 4th ed. (1986). MEAVE G. LEAKEY and RICHARD E. LEAKEY (eds.), *The Fossil Hominids and an Introduction to Their Context, 1968–1974* (1978), includes a lengthy catalog of the hominid finds at Koobi Fora. MICHAEL R. ZIMMERMAN and J. LAWRENCE ANGEL (eds.), *Dating and Age Determination of Biological Materials* (1986), includes discussion of dating human fossil remains. Two dated but still useful works are PHILLIP V. TOBIAS, *The Brain in Hominid Evolution* (1971), which contains data on endocranial capacity; and KENNETH P. OAKLEY, *Man the Tool-Maker,* 6th ed. (1972), an account of toolmaking development. PETER J. BOWLER, *Theories of Human Evolution: A Century of Debate, 1844–1944* (1986); and ROGER LEWIN, *Bones of Contention* (1987), discuss some of the major controversies in the field of paleoanthropology.

Good accounts of modern and fossil primates are found in J.R. NAPIER and P.H. NAPIER, *A Handbook of Living Primates: Morphology, Ecology, and Behaviour of Nonhuman Primates* (1967); W.E. Le GROS CLARK, *The Antecedents of Man: An Introduction to the Evolution of the Primates,* 3rd ed. (1971); ELWYN L. SIMONS, *Primate Evolution: An Introduction to Man's Place in Nature* (1972); ALISON JOLLY, *The Evolution of Primate Behavior,* 2nd ed. (1985); DARIS R. SWINDLER and J. ERWIN (eds.), *Comparative Primate Biology,* vol. 1, *Systematics, Evolution, and Anatomy* (1986); and JOHN G. FLEAGLE, *Primate Adaptation & Evolution* (1988). (F.C.H./M.H.D.)

Australopithecus: DONALD C. JOHANSON and MAITLAND A. EDEY, *Lucy: The Beginnings of Humankind* (1981), is a popular account of the discovery and interpretation of the Hadar hominids; and M.D. LEAKEY and J.M. HARRIS (eds.), *Laetoli: A Pliocene Site in Northern Tanzania* (1987), chronicles in detail the finds at this locality, including the tracks of hominid footprints. J.T. ROBINSON, *Early Hominid Posture and Locomotion* (1972), is also useful. FREDERICK E. GRINE, *The Evolutionary History of the "Robust" Australopithecines* (1989), is a comprehensive and specialized guide on the opinions about australopithecine variation. (B.Wo.)

Homo habilis: F.C. HOWELL, "Hominidae," in VINCENT J. MAGLIO and H.B.S. COOKE (eds.), *Evolution of African Mammals* (1978), pp. 154–248, provides thorough descriptions of the fossils assigned to *Homo habilis,* in addition to information about *Australopithecus.* RICHARD POTTS, *Early Hominid Activities at Olduvai* (1988), comments critically on hunting, scavenging, and social behaviour of the earliest humans, using the Olduvai archaeological evidence as a guide. P.V. TOBIAS, "The Brain of *Homo habilis:* A New Level of Organization in Cerebral Evolution," *Journal of Human Evolution,* 16(7–8):741–762 (1987), discusses the implications of brain anatomy for the use of language by the first representatives of *Homo.* C.B. STRINGER, "The Credibility of *Homo habilis,*" in BERNARD WOOD, LAWRENCE MARTIN, and PETER ANDREWS (eds.), *Major Topics in Primate and Human Evolution* (1986), pp. 266–294, advances the argument that fossils referred to as *Homo habilis* may actually document the presence of two distinct species in East Africa at about the same time.

Homo erectus: WILLIAM W. HOWELLS, "*Homo erectus*—Who, When, and Where: A Survey," *Yearbook of Physical Anthropology,* 23:1–23 (1980), is an excellent review. LARS-KÖNIG KÖNIGSSON (ed.), *Current Argument on Early Man* (1980), contains proceedings from a Nobel symposium that includes several papers on *H. erectus.* BECKY A. SIGMON and JEROME S. CYBULSKI (eds.), *Homo erectus* (1981), is a volume of essays by several authors who address history as well as current research on mid-Pleistocene humans. Useful works include RUKANG WU and SHENGLONG LIN, "Peking Man," *Scientific American,* 248(6):86–94 (June 1983), an account of excavations at Chou-k'ou-tien (Zhoukoudian), with information about the stone culture and hunting activities of Chinese *Homo erectus;* PETER ANDREWS and JENS LORENZ FRANZEN (eds.), *The Early Evolution of Man, with Special Emphasis on Southeast Asia and Africa* (1984), containing articles providing geologic background to the study of *Homo erectus* and discussion of anatomic features useful for defining the species; and G.P. RIGHTMIRE, "*Homo erectus* and Later Middle Pleistocene Humans," *Annual Review of Anthropology,* 17:239–260 (1988), a review emphasizing problems in the interpretation of the Middle Pleistocene fossil evidence. (G.P.Ri.)

Homo sapiens: (General): SEYMOUR W. ITZKOFF, *Triumph of the Intelligent: The Creation of Homo sapiens sapiens* (1985), which surveys the development of human intelligence and includes references to many classics in the field. FRED H. SMITH and FRANK SPENCER (eds.), *The Origins of Modern Humans: A World Survey of the Fossil Evidence* (1984), provides regional discussions of late archaic humans and the emergence of modern humans, with emphasis on the ancestor–descendant relationships between the regional groups. CAMERON D. OVEY (ed.), *The Swanscombe Skull: A Survey of Research on a Pleistocene Site* (1964), brings together the information that is available

concerning this fossil and its morphology, classification, dating, and site context. (M.H.D.)

(*Neanderthals*): ERIK TRINKAUS, "The Neandertals and Modern Human Origins," *Annual Review of Anthropology,* 15:193–218 (1986), provides an overview of behavioral contrasts between the Neanderthals and early modern humans, and his *The Shanidar Neandertals* (1983) explores the finds at Shanidar Cave in Iraq. GEORGE CONSTABLE, *The Neanderthals* (1973), is a useful introduction. ERIK TRINKAUS (ed.), *The Emergence of Modern Humans* (1990), discusses the human fossil and archaeological records as they relate to late archaic humans and the origin of modern humans. ERIK TRINKAUS and WILLIAM W. HOWELLS, "The Neanderthals," *Scientific American,* 241(6):118–133 (Dec. 1979), looks at the position of the Neanderthals in human ancestry and at their functional anatomy. (E.Tr.)

(*Cro-Magnons*): MARCELLIN BOULE and HENRI V. VALLOIS, *Fossil Men* (1957; originally published in French, 4th ed., 1952), is a well-illustrated general introduction to the field of fossil hominids, including an extended discussion of the Cro-Magnon peoples and the history of their discovery. *L'Homme de Cro-Magnon: anthropologie et archéologie* (1970), contains a series of papers by specialists in physical anthropology and prehistoric archaeology on the skeletal biology and culture of the Cro-Magnon peoples. ANDRÉ LEROI-GOURHAN, *Treasures of Prehistoric Art* (also published as *The Art of Prehistoric Man in Western Europe,* 1967; originally published in French, 1965), is a magnificently illustrated volume on the art of the Cro-Magnon peoples. HENRI V. VALLOIS and G. BILLY, "Nouvelles Recherches sur les hommes fossiles de l'abri de Cro-Magnon," *L'Anthropologie,* 69(1–2): 47–74 (1965) and 69(3–4): 249–272 (1965), is the most thorough study of the skeletal biology of the human remains from the Cro-Magnon shelter. E. GENET-VARCIN, *Les Hommes fossiles: découvertes et travaux depuis dix années* (1979), provides a descriptive inventory of Cro-Magnon discoveries throughout the world. (H.J.De.)

(*Asia and Australasia*): R.L. KIRK and A.G. THORNE (eds.), *The Origin of the Australians* (1976), contains essays on the nature and origins of the Aboriginal Australians and their relationship to both their nearest neighbours and the rest of the world. G.G. POPE, "Recent Advances in Far Eastern Paleoanthropology," *Annual Review of Anthropology,* 17:43–78 (1988); and RUKANG WU and JOHN W. OLSEN (eds.), *Palaeoanthropology and Palaeolithic Archaeology in the People's Republic of China* (1985), provide overviews of *H. sapiens* discoveries and research. *Atlas of Primitive Man in China* (1980), compiled by the Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Science, contains maps of fossil finds. (M.H.Wo.)

(*Africa*): Two essays in PAUL MELLARS and CHRISTOPHER STRINGER (eds.), *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans* (1989), are especially useful: GÜNTER BRAÜER, "The Evolution of Modern Humans: A Comparison of the African and Non-African Evidence," pp. 123–154; and J. DESMOND CLARK, "The Origins and Spread of Modern Humans: A Broad Perspective on the African Evidence," pp. 565–588. Other essays include G.P. RIGHTMIRE, "Africa and the Origins of Modern Humans," in RONALD SINGER and JOHN K. LUNDY (eds.), *Variation, Culture, and Evolution in African Populations* (1986), pp. 209–220; and C.B. STRINGER and P. ANDREWS, "Genetic and Fossil Evidence for the Origin of Modern Humans," *Science,* 239(4845):1263–68 (Mar. 11, 1988). (G.Br.)

Modern human populations: STANLEY M. GARN, *Human Races,* 3rd ed. (1971), offers a brief, lucid account of the process of race formation with a study of microevolution into small, subracial populations. FREDERICK S. HULSE, *The Human Species: An Introduction to Physical Anthropology,* 2nd ed. (1971), provides a general survey of human evolution and racial distribution. WILLIAM W. HOWELLS, *Mankind in the Making: The Story of Human Evolution,* rev. ed. (1967), also presents original ideas about modern races and their geographic distribution. A.E. MOURANT, ADA C. KOPÉC, and KAZIMIERA DOMANIEWSKA-SOBCZAK, *The Distribution of the Human Blood Groups and Other Polymorphisms,* 2nd ed. (1976), and a supplement with the same title (1983), are the classic works on human blood groups, their distribution, and racial significance. WILLIAM Z. RIPLEY, *The Races of Europe: A Sociological Study* (1899, reissued 1965), contains his original and often-quoted division of the European geographic race into Nordics, Alpines, and Mediterraneans. PETER E. NUTE and GEORGE STAMATOYANNOPOULOS, "Estimating Mutation Rates Using Abnormal Human Hemoglobins," *Yearbook of Physical Anthropology,* 27:135–151 (1984), estimates the rate of mutation of selected conditions. Studies of racial variation, trait inheritance, and continuing evolution include MARGARET MEAD *et al.* (eds.), *Science and the Concept of Race* (1968); CARL JAY BAJEMA, *Natural Selection in Human Populations: The Measurement of Ongoing Genetic Evolution in Contemporary Societies* (1971); JANE H. UNDERWOOD, *Human Variation and Human Microevolution* (1979); and STEPHEN MOLNAR, *Human Variation: Races, Types, and Ethnic Groups,* 2nd ed. (1983). (S.M.G.)

The Theory of Evolution

The diversity of the living world is staggering. More than 2,000,000 existing species of plants and animals have been named and described; many more remain to be discovered—from 10,000,000 to 30,000,000 according to some estimates. What is impressive is not just the numbers but also the incredible heterogeneity in size, shape, and way of life: from lowly bacteria, measuring less than one-thousandth of a millimetre in diameter, to the stately sequoias of California, rising 300 feet (100 metres) above the ground and weighing several thousand tons; from bacteria living in the hot springs of Yellowstone National Park at temperatures near the boiling point of water to fungi and algae thriving on the ice masses of Antarctica and in saline pools at −9° F (−23° C); and from the strange wormlike creatures discovered in dark ocean depths at thousands of feet below the surface to spiders and larkspur plants existing on Mt. Everest more than 19,868 feet above sea level.

The virtually infinite variations on life are the fruit of the evolutionary process. All living creatures are related by descent from common ancestors. Humans and other mammals are descended from shrewlike creatures that lived more than 150,000,000 years ago; mammals, birds, reptiles, amphibians, and fishes share as ancestors aquatic worms that lived 600,000,000 years ago; all plants and animals are derived from bacteria-like microorganisms that originated more than 3,000,000,000 years ago. Biological evolution is a process of descent with modification. Lineages of organisms change through generations; diversity arises because the lineages that descend from common ancestors diverge through time.

The 19th-century English naturalist Charles Darwin argued that organisms come about by evolution, and he provided a scientific explanation, essentially correct but incomplete, of how evolution occurs and why it is that organisms have features—such as wings, eyes, and kidneys—clearly structured to serve specific functions. Natural selection was the fundamental concept in his explanation. Genetics, a science born in the 20th century, reveals in detail how natural selection works and led to the development of the modern theory of evolution. Since the 1960s a related scientific discipline, molecular biology, has advanced enormously knowledge of biological evolution and has made it possible to investigate detailed problems that seemed completely out of reach a few years earlier—for example, how similar the genes of humans and chimpanzees might be (they differ in about 1 or 2 percent of the units that make up the genes).

This article discusses evolution as it applies generally to living things. For a discussion of human evolution, see the article EVOLUTION, HUMAN. For a more complete treatment of a discipline that has proved essential to the study of evolution, see GENETICS AND HEREDITY, THE PRINCIPLES OF. Specific aspects of evolution are discussed in the articles COLORATION, BIOLOGICAL; and MIMICRY. Applications of evolutionary theory to plant and animal breeding are discussed in the article FARMING AND AGRICULTURAL TECHNOLOGY. A detailed discussion of the life and thought of Charles Darwin is found in the article DARWIN.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 312.

The article is divided into the following sections:

History of evolutionary theory

EARLY IDEAS

All human cultures have developed their own explanations for the origin of the world, man, and other creatures. Traditional Judaism and Christianity explain the origin of living beings and their adaptations to their environments—wings, gills, hands, flowers—as the handiwork of an omniscient God. The philosophers of ancient Greece had their own creation myths. Anaximander proposed that animals could be transformed from one kind into another, and Empedocles speculated that they could be made up of various combinations of preexisting parts. Closer to modern evolutionary ideas were the proposals of early Church Fathers like Gregory of Nazianzus and Augustine, who maintained that not all species of plants and animals were created by God; rather, some had developed in historical times from God's creations. Their motivation was not biological but religious: it would have been impossible to hold representatives of all species in a single vessel such as Noah's ark; hence, some species must have come into existence only after the Noachian Flood.

The notion that organisms may change by natural processes was not investigated as a biological subject by Christian theologians of the Middle Ages, but it was, usually incidentally, considered as a possibility by many, includ-

ing Albertus Magnus and his student Thomas Aquinas. Aquinas concluded, after detailed discussion, that the development of living creatures like maggots and flies from nonliving matter like decaying meat was not incompatible with Christian faith or philosophy. But he left it to scientists to decide whether this actually happened in fact.

The idea of progress, particularly the belief in unbounded human progress, was central to the Enlightenment of the 18th century, particularly in France among such philosophers as Condorcet and Diderot and such scientists as Buffon. But belief in progress did not necessarily lead to the development of a theory of evolution. Pierre-Louis Moreau de Maupertuis proposed the spontaneous generation and extinction of organisms as part of his theory of origins, but he advanced no theory of evolution—*i.e.*, the transformation of one species into another through knowable, natural causes. Georges-Louis Leclerc, comte de Buffon, one of the greatest naturalists of the time, explicitly considered—and rejected—the possible descent of several species from a common ancestor. He postulated that organisms arise from organic molecules by spontaneous generation, so that there could be as many kinds of animals and plants as there are viable combinations of organic molecules.

The physician Erasmus Darwin, grandfather of Charles Darwin, offered in his *Zoonomia or the Laws of Organic Life* some evolutionary speculations, but they were not further developed and had no real influence on subsequent theories. The Swedish botanist Carolus Linnaeus devised the hierarchical system of plant and animal classification that is still in use in a modernized form. Although he insisted on the fixity of species, his classification system eventually contributed much to the acceptance of the concept of common descent.

Lamarck's theory of evolution

The great French naturalist Jean-Baptiste Lamarck held the enlightened view of his age that living organisms represent a progression, with humans as the highest form. From this idea he proposed, in the early years of the 19th century, the first broad theory of evolution. Organisms evolve through eons of time from lower to higher forms, a process still going on, always culminating in man. As organisms become adapted to their environments through their habits, modifications occur. Use of an organ or structure reinforces it; disuse leads to obliteration. The characteristics acquired by use and disuse, according to this theory, would be inherited. This assumption, later called the inheritance of acquired characteristics, was thoroughly disproved in the 20th century. Although his theory did not stand up in the light of later knowledge, Lamarck made important contributions to the gradual acceptance of biological evolution and stimulated countless later studies.

CHARLES DARWIN

The founder of the modern theory of evolution was Charles Darwin. The son and grandson of physicians, he enrolled as a medical student at the University of Edinburgh. After two years, however, he left to study at Cambridge University and prepare to become a clergyman. He was not an exceptional student, but he was deeply interested in natural history. On Dec. 27, 1831, a few months after his graduation from Cambridge, he sailed as a naturalist aboard the HMS *Beagle* on a round-the-world trip that lasted until October 1836. Darwin was often able to disembark for extended trips ashore to collect natural specimens.

The discovery of fossil bones from large extinct mammals in Argentina and the observation of numerous species of finches in the Galápagos Islands were among the events credited with stimulating Darwin's interest in how species originate. In 1859 he published *On the Origin of Species by Means of Natural Selection,* a treatise establishing the theory of evolution and, most important, the role of natural selection in determining its course. He published many other books as well, notably *The Descent of Man and Selection in Relation to Sex* (1871), which extends the theory of natural selection to human evolution.

Darwin must be seen as a great intellectual revolutionary who inaugurated a new era in the cultural history of mankind, an era that was the second and final stage of the Copernican revolution that had begun in the 16th and 17th centuries under the leadership of men such as Copernicus, Galileo, and Newton. The Copernican revolution marked the beginnings of modern science. Discoveries in astronomy and physics overturned traditional conceptions of the universe. The Earth was no longer seen as the centre of the universe but as a small planet revolving around one of a myriad of stars; the seasons, the rains that make crops grow, destructive storms, and other vagaries of weather all became understood as aspects of natural processes; the circumvolutions of the planets were now explained by simple laws that also accounted for the motion of projectiles on the Earth.

The universe explained by natural laws

The significance of these and other discoveries was that they led to a conception of the universe as a system of matter in motion governed by laws of nature. The workings of the universe no longer needed to be attributed to the ineffable will of the Creator, but were brought into the realm of science—an explanation of phenomena through natural laws. Physical phenomena like tides, eclipses, and positions of the planets could now be predicted whenever the causes were adequately known. Darwin accumulated evidence showing that evolution had occurred, that diverse organisms share common ancestors, and that living beings have changed drastically over the course of the Earth's history. More importantly, however, he extended to the living world the idea of nature as a system of matter in motion governed by natural laws.

Before Darwin, the origin of the Earth's living things, with their marvelous contrivances for adaptation, had been attributed to the design of an omniscient God. He had created the fish in the waters, the birds in the air, and all sorts of animals and plants on the land. God had endowed these creatures with gills for breathing, wings for flying, and eyes for seeing, and he had coloured birds and flowers so that man could enjoy them and recognize his wisdom. Christian theologians, from Thomas Aquinas on, had argued that the presence of design, so evident in living beings, demonstrates the existence of a supreme creator; the argument from design was Aquinas' "fifth way" for proving the existence of God. In 19th-century England, the eight Bridgewater Treatises were commissioned so that eminent scientists and philosophers would expand on the marvels of the natural world and thereby set forth "the Power, wisdom, and goodness of God as manifested in the Creation."

The British theologian William Paley in his *Natural Theology* (1802) used natural history, physiology, and other contemporary knowledge to elaborate the argument from design. If a person should find a watch, even in an uninhabited desert, Paley contended, the harmony of its many parts would force him to conclude that it had been created by a skilled watchmaker; and, Paley went on, how much more intricate and perfect in design is the human eye, with its transparent lens, its retina placed at the precise distance for forming a distinct image, and its large nerve transmitting signals to the brain.

The argument from design seems to be forceful. A ladder is made for climbing, a knife for cutting, and a watch for telling time; their functional design leads to the conclusion that they have been fashioned by a carpenter, a smith, or a watchmaker. Similarly, the obvious functional design of animals and plants seems to denote the work of a Creator. It was Darwin's genius that he provided a natural explanation for the organization and functional design of living beings.

Darwin accepted the facts of adaptation: hands are for grasping, eyes for seeing, lungs for breathing. But he showed that the multiplicity of plants and animals, with their exquisite and varied adaptations, could be explained by a process of natural selection, without recourse to a Creator or any designer agent. He brought the living world into the realm of natural science, thereby completing the Copernican revolution. All natural phenomena were henceforth opened to explanation by natural causes and viewed as the result of physical processes governed by natural laws. This achievement of Darwin's had intellectual and cultural implications more profound and lasting than his multipronged evidence that convinced contemporaries of the fact of evolution.

Darwin's theory of natural selection

Darwin's theory of natural selection is summarized in the *Origin of Species* as follows:

> As many more individuals are produced than can possibly survive, there must in every case be a struggle for existence, either one individual with another of the same species, or with the individuals of distinct species, or with the physical conditions of life.... Can it, then, be thought improbable, seeing that variations useful to man have undoubtedly occurred, that other variations useful in some way to each being in the great and complex battle of life, should sometimes occur in the course of thousands of generations? If such do occur, can we doubt (remembering that many more individuals are born than can possibly survive) that individuals having any advantage, however slight, over others, would have the best chance of surviving and of procreating their kind? On the other hand, we may feel sure that any variation in the least degree injurious would be rigidly destroyed. This preservation of favourable variations and the rejection of injurious variations, I call Natural Selection.

Natural selection was proposed by Darwin primarily to account for the adaptive organization of living beings; it is a process that promotes or maintains adaptation. Evolutionary change through time and evolutionary diversification (multiplication of species) are not directly promoted by natural selection, but they often ensue as by-products of natural selection as it fosters adaptation to different environments. (See below *The process of evolution: The concept of natural selection.*)

MODERN CONCEPTIONS

The Darwinian aftermath. The publication of the *Origin of Species* produced considerable public excitement. Scientists, politicians, clergymen, and notables of all kinds read and discussed the book, defending or deriding Darwin's ideas. The most visible actor in the controversies immediately following publication was T.H. Huxley, known as "Darwin's bulldog," who defended the theory of evolution with articulate and sometimes mordant words on public occasions as well as in numerous writings. Evolution by natural selection was indeed a favourite topic in society salons during the 1860s and beyond. But serious scientific controversies also arose, first in Britain and then on the Continent and in the United States.

Collaboration of Darwin and Wallace

One occasional participant in the discussion was the naturalist Alfred Russel Wallace, who had hit upon the idea of natural selection independently and had sent a short manuscript to Darwin from the Malay archipelago. On July 1, 1858, one year before the publication of the *Origin,* a paper jointly written by Wallace and Darwin was presented, in the absence of both, to the Linnean Society in London—with apparently little notice. Greater credit is duly given to Darwin than to Wallace for the idea of evolution by natural selection; Darwin developed the theory in considerably more detail, provided far more evidence for it, and was primarily responsible for its acceptance. Wallace's views differed from Darwin's in several ways, most importantly in that Wallace did not think natural selection sufficient to account for the origin of man, which in his view required direct divine intervention.

A younger contemporary of Darwin, with considerable influence during the latter part of the 19th and early 20th centuries, was Herbert Spencer. He was a philosopher rather than a biologist, but he became an energetic proponent of evolutionary ideas, popularized a number of slogans, like "survival of the fittest" (which was taken up by Darwin in later editions of the *Origin*), and engaged in social and metaphysical speculations. His ideas considerably damaged proper understanding and acceptance of the theory of evolution by natural selection. Darwin wrote of Spencer's speculations:

> His deductive manner of treating any subject is wholly opposed to my frame of mind.... His fundamental generalizations (which have been compared in importance by some persons with Newton's laws!) which I dare say may be very valuable under a philosophical point of view, are of such a nature that they do not seem to me to be of any strictly scientific use.

Most pernicious was the crude extension of the notion of "struggle for existence" to human economic and social life that became known as social Darwinism.

The most serious difficulty facing Darwin's evolutionary theory was the lack of an adequate theory of inheritance that would account for the preservation through the generations of the variations on which natural selection was supposed to act. Current theories of "blending inheritance" proposed that offspring merely struck an average between the characteristics of their parents. As Darwin became aware, blending inheritance (including his own theory of "pangenesis") could not account for the conservation of variations, because differences among variant offspring would be halved each generation, rapidly reducing the original variation to the average of the preexisting characteristics.

The missing link in Darwin's argument was provided by Mendelian genetics. About the time the *Origin of Species* was published, the Augustinian monk Gregor Mendel was starting a long series of experiments with peas in the garden of his monastery in Brünn, Austria-Hungary (now Brno, Czech.). These experiments and the analysis of their results are by any standard an example of masterly scientific method. Mendel's paper, published in 1866 in the *Proceedings* of the Natural Science Society of Brünn, formulated the fundamental principles of the theory of heredity that is still current. His theory accounts for biological inheritance through particulate factors (genes) inherited one from each parent, which do not mix or blend but segregate in the formation of the sex cells, or gametes.

Mendel's principles of heredity

Mendel's discoveries, however, remained unknown to Darwin and, indeed, did not become generally known until 1900, when they were simultaneously rediscovered by a number of scientists on the Continent. In the meantime Darwinism, in the latter part of the 19th century, faced an alternative evolutionary theory known as neo-Lamarckism. This hypothesis shared with Lamarck's the importance of use and disuse in the development and obliteration of organs, and it added the notion that the environment acts directly on organic structures, which explained their adaptation to the way of life and environment of the organism. Adherents of this theory discarded natural selection as an explanation for adaptation to the environment.

Prominent among the defenders of natural selection was the German biologist August Weismann, who in the 1880s published his germ-plasm theory. He distinguished two substances that make up an organism: the soma, which comprises most body parts and organs, and the germ plasm, which contains the cells that give rise to the gametes and hence to progeny. Early in the development of an egg, the germ plasm becomes segregated from the soma; that is, from the cells that give rise to the rest of the body. This notion of a radical separation between germ and soma prompted Weismann to assert that inheritance of acquired characteristics was impossible, and it opened the way for his championship of natural selection as the only major process that would account for biological evolution. Weismann's ideas became known after 1896 as neo-Darwinism.

The synthetic theory. The rediscovery in 1900 of Mendel's theory of heredity, by Hugo de Vries of The Netherlands and others, led to an emphasis on the role of heredity in evolution. De Vries proposed a new theory of evolution known as mutationism, which essentially did away with natural selection as a major evolutionary process. According to de Vries (joined by other geneticists such as William Bateson in England), there are two kinds of variation that take place in organisms. One is the "ordinary" variability observed among individuals of a species, which is of no lasting consequence in evolution because, according to de Vries, it could not "lead to a transgression of the species border even under conditions of the most stringent and continued selection." The other consists of the changes brought about by mutations, spontaneous alterations of genes that yield large modifications of the organism and gave rise to new species: "The new species thus originates suddenly, it is produced by the existing one without any visible preparation and without transition."

Mutationists and biometricians

Mutationism was opposed by many naturalists, and in particular by the so-called biometricians, led by Karl Pearson, who defended Darwinian natural selection as the

major cause of evolution through the cumulative effects of small, continuous, individual variations (which the biometricians assumed passed from one generation to the next without being limited by Mendel's laws of inheritance).

The controversy between mutationists (also referred to at the time as Mendelians) and biometricians approached a resolution in the 1920s and '30s through the theoretical work of geneticists. They used mathematical arguments to show, first, that continuous variation (in such characteristics as size, number of eggs laid, and the like) could be explained by Mendel's laws; and second, that natural selection acting cumulatively on small variations could yield major evolutionary changes in form and function. Distinguished members of this group of theoretical geneticists were R.A. Fisher and J.B.S. Haldane in Britain and Sewall Wright in the United States. Their work contributed to the downfall of mutationism and, most importantly, provided a theoretical framework for the integration of genetics into Darwin's theory of natural selection. Yet their work had a limited impact on contemporary biologists because it was formulated in a mathematical language that most of them could not understand; because it was almost exclusively theoretical, with little empirical corroboration; and because it was limited in scope, largely omitting many issues, like speciation, that were of great importance to evolutionists.

A major breakthrough came in 1937 with the publication of *Genetics and the Origin of Species* by Theodosius Dobzhansky, a Russian-born American naturalist and experimental geneticist. Dobzhansky's book advanced a reasonably comprehensive account of the evolutionary process in genetic terms, laced with experimental evidence supporting the theoretical argument. *Genetics and the Origin of Species* may be considered the most important landmark in the formulation of what came to be known as the synthetic theory of evolution, effectively combining Darwinian natural selection and Mendelian genetics. It had an enormous impact on naturalists and experimental biologists, who rapidly embraced the new understanding of the evolutionary process as one of genetic change in populations. Interest in evolutionary studies was greatly stimulated, and contributions to the theory soon began to follow, extending the synthesis of genetics and natural selection to a variety of biological fields.

Influence of Dobzhansky

The main writers who, together with Dobzhansky, may be considered the architects of the synthetic theory were the zoologists Ernst Mayr and Sir Julian Huxley, the paleontologist George G. Simpson, and the botanist George Ledyard Stebbins. These researchers contributed to a burst of evolutionary studies in the traditional biological disciplines and in some emerging ones—notably population genetics and, later, evolutionary ecology. By 1950 acceptance of Darwin's theory of evolution by natural selection was universal among biologists, and the synthetic theory had become widely adopted.

Later developments. The most important line of investigation since 1950 has been the application of molecular biology to evolutionary studies. In 1953 James Watson and Francis Crick deduced the structure of DNA (deoxyribonucleic acid), the hereditary material contained in the chromosomes of every cell's nucleus. The genetic information is contained within the sequence of nucleotides that make up the chainlike DNA molecules. This information determines the sequence of amino acids in the proteins, including enzymes responsible for the organism's fundamental life processes. Genetic information contained in the DNA can thus be investigated by examining the sequences of amino acids in the proteins.

In the mid-1960s techniques like electrophoresis and selective assay of enzymes became available for the rapid and inexpensive study of differences among enzymes and other proteins. The application of these techniques to evolutionary problems made possible the pursuit of issues that earlier could not be investigated; for example, exploring the extent of genetic variation in natural populations (which sets bounds to their evolutionary potential) and determining the amount of genetic change that occurs during the formation of new species.

Comparisons of the amino acid sequences of proteins in different species provided quantitatively precise measures of species divergence, a considerable improvement over the typically qualitative evaluations obtained by comparative anatomy and other evolutionary subdisciplines. In 1968 the Japanese geneticist Motoo Kimura proposed the neutrality theory of molecular evolution, which assumes that at the level of DNA and protein sequence many changes are adaptively neutral and have little or no effect on the molecule's function. If the neutrality theory is correct, there should be a "molecular clock" of evolution; that is, the degree of divergence between species in amino acid or nucleotide sequence would provide a reliable estimate of the time since their divergence. This would make possible a reconstruction of evolutionary history that would reveal the order of branching of different lineages, such as those leading to humans, chimpanzees, and orangutans, as well as the time in the past when the lineages split from one another.

The "molecular clock"

During the 1970s and '80s it gradually became clear that the molecular clock is not exact; nevertheless, it has continued to provide the most reliable source of evidence for reconstructing a history of evolution. The techniques of DNA cloning and sequencing have provided a new and more powerful means of investigating evolution at the molecular level. The fruits of this new development began to accumulate during the 1980s.

The earth sciences have also experienced, in the second half of the 20th century, a conceptual revolution with considerable consequence to the study of evolution. The science of plate tectonics has shown that the configuration and position of the continents and oceans are dynamic, rather than static, features of the Earth. Oceans grow and shrink, while continents break into fragments or coalesce into larger masses. The continents move across the Earth's surface at rates of a few centimetres a year, and over millions of years of geologic history this profoundly alters the face of the Earth, causing major climatic changes along the way. These previously unsuspected massive modifications of the planet's environments have of necessity been reflected in the evolutionary history of life. Biogeography, the evolutionary study of plant and animal distribution, has been revolutionized by the knowledge, for example, that Africa and South America were part of a single landmass some 200,000,000 years ago and that the Indian subcontinent was not connected with Asia until recent geologic times.

Ecology, the study of the interactions of organisms with their environments, has evolved from descriptive studies—"natural history"—into a vigorous biological discipline with a strong mathematical component, both in the development of theoretical models and in the collection and analysis of quantitative data. Evolutionary ecology is an active field of evolutionary biology; another is evolutionary ethology, the study of animal behaviour. Sociobiology, the evolutionary study of social behaviour, is perhaps the most active subfield of ethology. It is also the most controversial, because of its extension to human societies.

IMPACT AND ACCEPTANCE OF EVOLUTIONARY THEORY

The theory of evolution makes statements about three different, though related, issues: (1) the fact of evolution; that is, that organisms are related by common descent; (2) evolutionary history—the details of when lineages split from one another and of the changes that occurred in each lineage; and (3) the mechanisms or processes by which evolutionary change occurs.

The first issue is the most fundamental and the one established with utmost certainty. Darwin gathered much evidence in its support, but the evidence has accumulated continuously ever since, derived from all biological disciplines. The evolutionary origin of organisms is today a scientific conclusion established with the kind of certainty attributable to such scientific concepts as the roundness of the Earth, the motions of the planets, and the molecular composition of matter. This degree of certainty beyond reasonable doubt is what is implied when biologists say that evolution is a "fact"; the evolutionary origin of organisms is accepted by virtually every biologist.

The scientific fact of evolution

But the theory of evolution goes much beyond this first

issue, the general affirmation that organisms evolve. The second and third issues involve seeking to ascertain the evolutionary relationships between particular organisms and the events of evolutionary history, as well as to explain how and why evolution takes place. These are matters of active scientific investigation. Some conclusions are well established; for example, that the chimpanzee and gorilla are more closely related to humans than is any of those three species to the baboon or other monkeys; or that natural selection, the process postulated by Darwin, explains the adaptive configuration of such features as the human eye and the wings of birds. Many matters are less certain, others are conjectural, and still others—such as the characteristics of the first living things and when they came about—remain completely unknown.

Since Darwin, the theory of evolution has gradually extended its influence to other biological disciplines, from physiology to ecology and from biochemistry to systematics. All biological knowledge now includes the phenomenon of evolution. In the words of Dobzhansky, "Nothing in biology makes sense except in the light of evolution."

The term evolution and the concept of change through time have also become incorporated into scientific language well beyond biology, and even into common language. Astrophysicists speak of the evolution of the solar system or of the universe; geologists, of the evolution of the Earth's mantle; psychologists, of the evolution of the mind; anthropologists, of the evolution of cultures; art historians, of the evolution of architectural styles; and couturiers, of the evolution of fashion. These and other disciplines share only the slightest commonality of meaning: the notion of gradual, and perhaps directional, change over the course of time.

Darwin's notion of natural selection has also been extended to other areas of human discourse, particularly in the fields of sociopolitical theory and economics. The extension can only be metaphorical, because in Darwin's intended meaning natural selection applies only to hereditary variations in entities endowed with biological reproduction, that is, to living organisms. That natural selection is a natural process in the living world has been taken by some as a justification for ruthless competition and for "survival of the fittest" in the struggle for economic advantage or for political hegemony. Social Darwinism was an influential social philosophy in some circles through the late 19th and early 20th centuries. At the other end of the political spectrum, Marxist theorists have resorted to evolution by natural selection as an explanation for mankind's political history.

These uses and abuses of the terms evolution and natural selection have in turn stimulated resistance against biological evolution and natural selection. In addition, the theory of evolution has been seen by some people as incompatible with religious beliefs, particularly those of Christianity.

Christianity and evolutionism

The first chapters of the book of Genesis describe God's creation of the world, the plants, the animals, and man. A literal interpretation of Genesis seems incompatible with the gradual evolution of humans and other organisms by natural processes. Independently of the biblical narrative, the Christian beliefs in the immortality of the soul and in man as "created in the image of God" have appeared to many as contrary to the evolutionary origin of man from nonhuman animals.

Religiously motivated attacks started during Darwin's lifetime. In 1874 Charles Hodge, an American Protestant theologian, published *What Is Darwinism?,* one of the most articulate assaults on evolutionism. Hodge perceived Darwin's theory as "the most thoroughly naturalistic that can be imagined and far more atheistic than that of his predecessor Lamarck." He argued that the design of the human eye evinces that "it has been planned by the Creator, like the design of a watch evinces a watchmaker." He concluded that "the denial of design in nature is actually the denial of God."

Other Protestant theologians saw a solution to the difficulty in the idea that God operates through intermediate causes. The origin and motion of the planets could be explained by the law of gravity and other natural processes without denying God's creation and providence. Similarly, evolution could be seen as the natural process through which God brought living beings into existence and developed them according to his plan. Thus, A.H. Strong, the president of Rochester (N.Y.) Theological Seminary, wrote in his *Systematic Theology* (1885): "We grant the principle of evolution, but we regard it as only the method of divine intelligence." The brutish ancestry of man was not incompatible with his excelling status as a creature in the image of God. Strong drew an analogy with Christ's miraculous conversion of water into wine: "The wine in the miracle was not water because water had been used in the making of it, nor is man a brute because the brute has made some contributions to its creation."

Arguments for and against Darwin's theory came from Roman Catholic theologians as well. Gradually, well into the 20th century, evolution by natural selection came to be accepted by the majority of Christian writers. Pope Pius XII in his encyclical *Humani Generis* (1950; "Of the Human Race") acknowledged that biological evolution was compatible with the Christian faith, although he argued that God's intervention was necessary for the creation of the human soul. In 1981 Pope John Paul II stated in an address to the Pontifical Academy of Sciences:

> The Bible itself speaks to us of the origin of the universe and its make-up, not in order to provide us with a scientific treatise but in order to state the correct relationships of man with God and with the universe. Sacred scripture wishes simply to declare that the world was created by God, and in order to teach this truth it expresses itself in the terms of the cosmology in use at the time of the writer.... Any other teaching about the origin and make-up of the universe is alien to the intentions of the Bible, which does not wish to teach how the heavens were made but how one goes to heaven.

The Pope's point was that it would be a blunder to mistake the Bible for an elementary book of astronomy, geology, and biology. The argument was clearly directed against Christian Fundamentalists who see in Genesis a literal description of how the world was created by God. Biblical Fundamentalists make up a minority of Christians, but they have periodically gained considerable public and political influence in the United States. During the decade of the 1920s, more than 20 state legislatures were influenced by them to debate antievolution laws, and four states—Arkansas, Mississippi, Oklahoma, and Tennessee—prohibited the teaching of evolution in their public schools. A spokesman for the antievolutionists was William Jennings Bryan, three times the unsuccessful Democratic candidate for the presidency, who said in 1922, "We will drive Darwinism from our schools." In 1925 Bryan took part in the prosecution of John T. Scopes, a high school teacher in Dayton, Tenn., who had admittedly violated the state's law forbidding the teaching of evolution.

The reaction of Christian Fundamentalists

In 1968 the Supreme Court of the United States declared unconstitutional any law banning the teaching of evolution in public schools. Since that time Christian Fundamentalists have introduced bills in a number of state legislatures ordering that the teaching of "evolution-science" be balanced by allocating equal time to "creation-science." Creation-science maintains that all kinds of organisms abruptly came into existence at the Creation, that the world is only a few thousand years old, and that the Noachian Flood was an actual event which only one pair of each animal species survived. In the 1980s Arkansas and Louisiana passed acts requiring the balanced treatment of evolution-science and creation-science in the schools, but opponents successfully challenged the acts as violations of the constitutionally mandated separation of church and state.

The evidence for evolution

Darwin and other 19th-century biologists found compelling evidence for biological evolution in the comparative study of living organisms, in their geographic distribution, and in the fossil remains of extinct organisms. In the 20th century the evidence from these sources has become considerably stronger and more comprehensive. New biological disciplines—genetics, biochemistry, phys-

Sources of information

iology, ecology—have supplied powerful additional evidence. Molecular biology, the most recent and successful of all biological disciplines, furnishes extensive, consistent, and detailed confirmation of evolution. The amount of information about evolutionary history stored in the DNA and proteins of living things is virtually unlimited. Only a want of resources now precludes the reconstruction of every detail of the phylogenetic history of life on Earth.

Evolutionists are no longer concerned with obtaining evidence to support the fact of evolution, but rather with what sorts of knowledge can be obtained from different sources of evidence. The following sections identify the most productive of these sources and illustrate the types of information they have provided.

THE FOSSIL RECORD

Paleontologists have recovered and studied the fossil remains of many thousands of organisms that lived in the past. The fossil record shows that many kinds of extinct organisms were very different in form from any now living. It also shows successions of organisms through time, manifesting the transition from one form to another.

When an organism dies, it is usually destroyed by other forms of life and by weathering processes. On rare occasions, some body parts—particularly hard ones like shells, teeth, or bones—are preserved by being buried in mud or protected in some other way from predators and weather. Eventually they may become petrified and preserved indefinitely with the rocks in which they are embedded. Methods such as measuring radioactive decay make it possible to estimate the time period when the rocks, and the fossils associated with them, were formed.

Radioactive dating indicates that the Earth was formed about 4,500,000,000 years ago. The earliest fossils resemble microorganisms such as bacteria and blue-green algae; the oldest ones appear in rocks 3,500,000,000 old. The oldest animal fossils, about 700,000,000 years old, come from small wormlike creatures with soft bodies. Numerous fossils belonging to many living phyla and exhibiting mineralized skeletons appear in rocks about 570,000,000 years old. These organisms are different from organisms living now and from those living at intervening times. Some are so radically different that paleontologists have created new phyla in order to classify them. The first vertebrates, animals with backbones, appeared about 400,000,000 years ago; the first mammals less than 200,000,000 years ago. The history of life recorded by fossils presents compelling evidence of evolution.

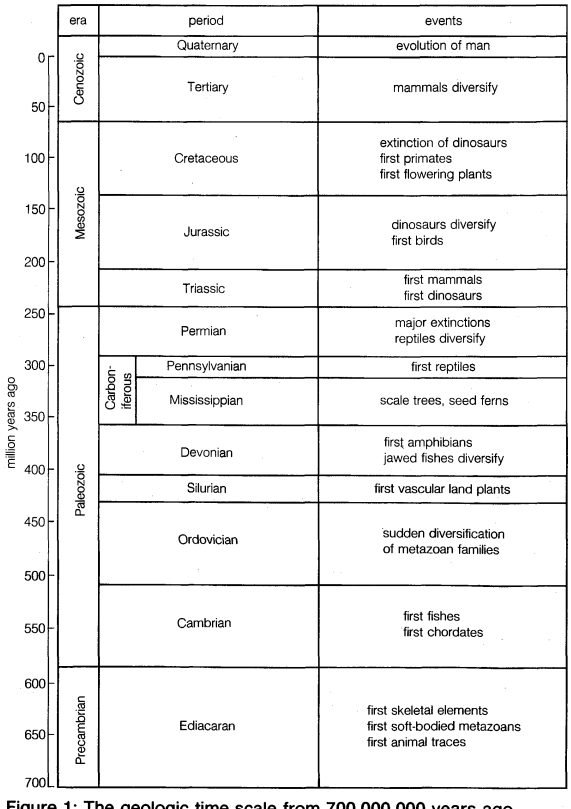


Figure 1: The geologic time scale from 700,000,000 years ago to the present, showing major evolutionary events.

The fossil record is incomplete. Of the small proportion of organisms preserved as fossils, only a tiny fraction have been recovered and studied by paleontologists. But the succession of forms over time has been in some cases reconstructed in detail. One example is the evolution of the horse (see Figure 2). It began with the dawn horse

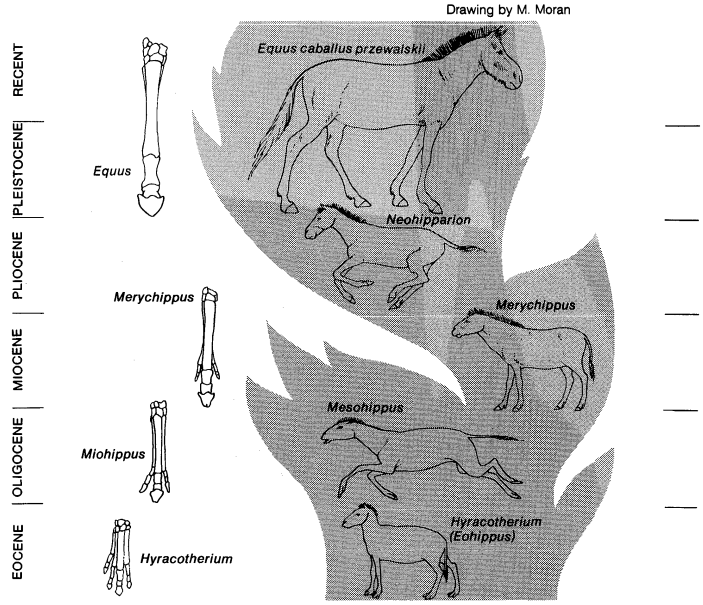


Figure 2: Evolution of the horse.

(genus *Hyracotherium*), an animal the size of a dog, with several toes on each foot and dentition appropriate for browsing, which evolved over 50,000,000 years ago; the most recent form is *Equus*, the modern horse, much larger in size, one-toed, and with teeth appropriate for grazing. The transitional forms are well preserved as fossils, as are many other kinds of extinct horses that evolved in different directions and left no living descendants.

Paleontologists have also been able to recover and reconstruct radical transitions in form and function. The lower jaw of reptiles contains several bones, that of mammals only one; the other bones in the reptile jaw evolved into bones now found in the mammalian ear. This would seem an unlikely transition. A bone being either in the jaw or in the ear, it is hard to imagine what function it could have during the intermediate stages. Yet paleontologists discovered two transitional forms of therapsids (mammal-like reptiles) with a double jaw joint—one joint consisting of the bones that persist in the mammalian jaw, the other composed of the quadrate and articular bones, which eventually became the hammer and anvil of the mammalian ear.

The "missing link"

For skeptical contemporaries of Darwin, the "missing link"—the absence of any transitional form between apes and humans—was a battle cry, as it remained for uninformed people afterward. Not one but many creatures intermediate between living apes and humans have since been found as fossils. *Australopithecus,* a hominid that lived 3,000,000 or 4,000,000 years ago, had an upright human stance but a cranial capacity of less than 500 cubic centimetres—comparable to that of a gorilla or chimpanzee and just about one-third that of humans. Its head displayed an odd mixture of ape and human characteristics: a low forehead and a long, ape-like face, but with teeth proportioned like those of humans. Along with increased cranial capacity, other human characteristics have been found in *Homo habilis,* which lived about 1,500,000 to 2,000,000 years ago and had a cranial capacity of more than 600 cubic centimetres, and *Homo erectus,* which lived between 500,000 and more than 1,000,000

years ago and had a cranial capacity of 800 to 1,100 cubic centimetres.

STRUCTURAL SIMILARITIES

The skeletons of turtles, horses, humans, birds, and bats are strikingly similar, in spite of the different ways of life of these animals and the diversity of their environments. The correspondence, bone by bone, can easily be seen in the limbs (Figure 3), but also in every other part of the body. From a purely practical point of view, it is incomprehensible that a turtle should swim, a horse run, a person write, and a bird or bat fly with structures built of the same bones. An engineer could design better limbs in each case. But if it is accepted that all of these skeletons inherited their structures from a common ancestor and became modified only as they adapted to different ways of life, the similarity of their structures makes sense.

Comparative anatomy investigates the homologies, or inherited similarities, among organisms in bone structure and in other parts of the body. The correspondence of structures is typically very close among some organisms—the different varieties of songbirds, for instance—but becomes less so as the organisms compared are less closely related in their evolutionary history. The similarities are less detailed between mammals and birds than they are among mammals, and less yet between mammals and fishes. Similarities in structure, therefore, not only manifest evolution but also help to reconstruct the phylogeny, or evolutionary history, of organisms.

An explanation of why most organismic structures are not perfect is also revealed by comparative anatomy. Like the forelimbs of turtles, horses, humans, birds, and bats, an organism's body parts are less than perfectly adapted because they are modified from an inherited structure rather than designed from completely "raw" materials for a specific purpose. The imperfection of structures is evidence for evolution and against design.

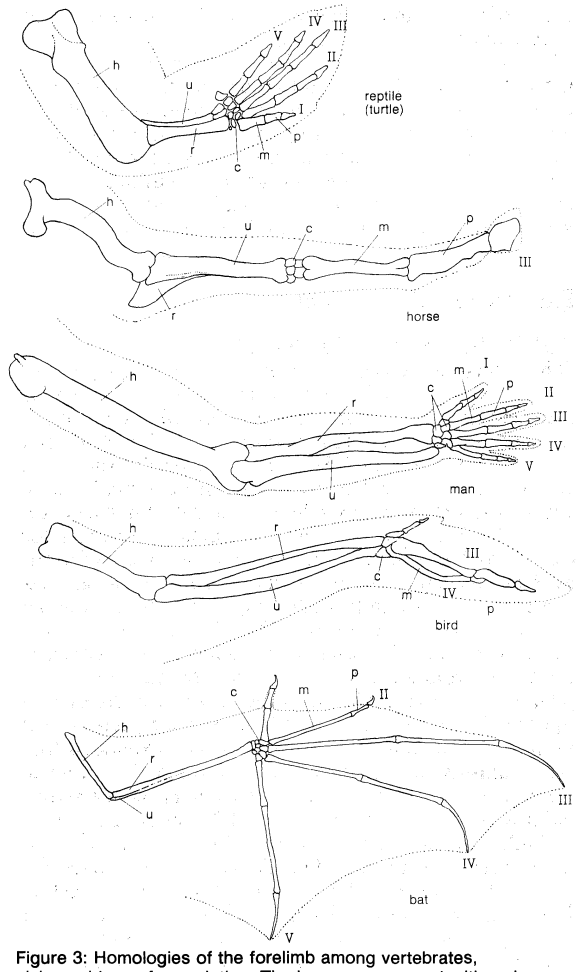


Figure 3: Homologies of the forelimb among vertebrates, giving evidence for evolution. The bones correspond, although they are adapted to the specific mode of life of the animal. The abbreviations are: h, humerus; r, radius; u, ulna; c, carpals; m, metacarpal; p, phalanx. The Roman numerals indicate corresponding digits.

EMBRYONIC DEVELOPMENT AND VESTIGES

Darwin and his followers found support for evolution in the comparative study of embryology—the science that investigates the development of organisms from fertilized egg to time of birth or hatching. Vertebrates, from fishes through lizards to humans, develop in ways that are remarkably similar during early stages, but they become more and more differentiated as the embryos approach maturity. The similarities persist longer between organisms that are more closely related (man and monkey) than between those less closely related (man and shark). Common developmental patterns reflect evolutionary kinship. Lizards and humans share developmental patterns inherited from their remote common ancestor; the inherited pattern was modified only as the separate descendant lineages evolved in different directions. The common embryonic stages of the two creatures reflect the constraints imposed by this common inheritance, which prevents changes that have not been necessitated by their diverging environment and way of life.

Similarity of vertebrate embryos

Human and other nonaquatic embryos exhibit gill slits even though they never breathe through gills. These slits are found in the embryos of all vertebrates because they share as common ancestors the fish in which these structures first evolved. Human embryos also exhibit by the fourth week of development a well-defined tail, which reaches maximum length when the embryo is six weeks old. Similar embryonic tails are found in other mammals, such as dogs, horses, and monkeys; in humans, however, the tail eventually shortens, persisting only as a rudiment in the adult coccyx.

A close evolutionary relationship between organisms that appear drastically different as adults can sometimes be recognized by their embryonic homologies. Barnacles are sedentary crustaceans with little apparent likeness to such crustaceans as lobsters, shrimps, or copepods. Yet barnacles pass through a free-swimming larval stage, the nauplius, which is unmistakably similar to other crustacean larvae.

Embryonic rudiments that never fully develop, such as the gill slits in humans, are common in all sorts of animals. Some, however, like the tail rudiment in humans, persist as adult vestiges reflecting evolutionary ancestry. The most familiar rudimentary organ in humans is the vermiform appendix. This wormlike structure attaches to a short section of intestine called the cecum, which is located at the point where the large and small intestines join. The human vermiform appendix is a functionless vestige of a fully developed organ present in other mammals, such as the rabbit and other herbivores, where a large cecum and appendix store vegetable cellulose to enable its digestion with the help of bacteria. Vestiges are instances of imperfections that argue against creation by design but are fully understandable as a result of evolution.

BIOGEOGRAPHY

Darwin also saw a confirmation of evolution in the geographic distribution of plants and animals, and later knowledge has reinforced his observations. For example, there are about 1,500 species of *Drosophila* vinegar flies in the world; nearly one-third of them live in Hawaii and nowhere else, although the total area of the archipelago is less than one-twentieth the area of California. There are also in Hawaii more than 1,000 species of snails and other land mollusks that exist nowhere else. This unusual diversity is easily explained by evolution. The Hawaiian Islands are extremely isolated and have had few colonizers; those species that arrived there found many unoccupied ecological niches, or local environments suited to sustain them and lacking predators that would prevent them from multiplying. In response, they rapidly diversified; this process of diversifying in order to fill in ecological niches is called adaptive radiation.

Evolutionary significance of species distribution

The continents of the world have their own distinc-

tive fauna and flora. In Africa there are rhinoceroses, hippopotamuses, lions, hyenas, giraffes, zebras, lemurs, monkeys with narrow noses and nonprehensile tails, chimpanzees, and gorillas. South America, which extends over much the same latitudes as Africa, has none of these animals but has different ones: pumas, jaguars, tapirs, llamas, raccoons, opossums, armadillos, and monkeys with broad noses and large prehensile tails.

These vagaries of biogeography are not due solely to the suitability of the different environments. There is no reason to believe that South American animals are not well suited to live in Africa or those of Africa to live in South America. The Hawaiian Islands are no better suited than other Pacific islands for *Drosophila* flies, nor are they less hospitable than other parts of the world for many absent organisms. In fact, although no large mammals are native to the islands, pigs and goats have multiplied there as wild animals since being introduced by humans. This absence of many species from a hospitable environment in which an extraordinary variety of other species flourishes can be explained by the theory of evolution, which holds that species can exist and evolve only in geographic areas that were colonized by their ancestors.

MOLECULAR BIOLOGY

The field of molecular biology has emerged during the mid-20th century. This new discipline has unveiled the nature of hereditary material and the workings of organisms at the level of enzymes and other molecules. Molecular biology provides the most detailed and convincing evidence available for biological evolution.

It is now known that the hereditary material, DNA, and the enzymes that govern all life processes hold information about an organism's ancestry. This information has made it possible to reconstruct evolutionary events that were previously unknown and to confirm and adjust the view of events that already were known. The precision with which events of evolution can be reconstructed is one reason the evidence from molecular biology is so compelling. Another reason is that molecular evolution has shown all living organisms, from bacteria to humans, to be related by descent from common ancestors.

Molecular uniformity of all organisms

A remarkable uniformity exists in the molecular components of organisms—in the nature of the components as well as in the ways in which they are assembled and used. In all bacteria, plants, animals, and humans, the DNA comprises a different sequence of the same four component nucleotides, and all of the various proteins are synthesized from different combinations and sequences of the same 20 amino acids, although several hundred other amino acids do exist. The genetic "code" by which the information contained in the nuclear DNA is passed on to proteins is everywhere the same. Similar metabolic pathways are used by the most diverse organisms to produce energy and to make up the cell components.

This unity reveals the genetic continuity and common ancestry of all organisms. There is no other rational way to account for their molecular uniformity when numerous alternative structures are equally likely. The genetic code may serve as an example. Each particular sequence of three nucleotides in the nuclear DNA acts as a pattern, or code, for the production of exactly the same amino acid in all organisms. This is no more necessary than it is for a language to use a particular combination of letters to represent a particular reality. If it is found that certain sequences of letters—*planet, tree, woman*—are used with identical meanings in a number of different books, one can be sure that the languages used in those books are of common origin.

Genes and proteins are long molecules that contain information in the sequence of their components in much the same way as sentences of the English language contain information in the sequence of their letters and words. The sequences that make up the genes are passed on from parents to offspring, identical except for occasional changes introduced by mutations. To illustrate, assume that two books are being compared; both books are 200 pages long and contain the same number of chapters. Closer examination reveals that the two books are identical page for page and word for word, except that an occasional word—say one in 100—is different. The two books cannot have been written independently; either one has been copied from the other or both have been copied, directly or indirectly, from the same original book. Similarly, if each nucleotide is represented by one letter, the complete sequence of nucleotides in the DNA of a higher organism would require several hundred books of hundreds of pages, with several thousand letters on each page. When the "pages" (or sequence of nucleotides) in these "books" (organisms) are examined one by one, the correspondence in the "letters" (nucleotides) gives unmistakable evidence of common origin.

The arguments presented above are based on different grounds, although both attest to evolution. Using the alphabet analogy, the first argument says that languages that use the same dictionary—the same genetic code and the same 20 amino acids—cannot be of independent origin. The second argument, concerning similarity in the sequence of nucleotides in the DNA or the sequence of amino acids in the proteins, says that books with very similar texts cannot be of independent origin.

The evidence of evolution revealed by molecular biology goes one step further. The degree of similarity in the sequence of nucleotides or of amino acids can be precisely quantified. For example, cytochrome c (a protein molecule) of humans and chimpanzees consists of the same 104 amino acids in exactly the same order; but differs from that of rhesus monkeys by one amino acid, that of horses by 11 additional amino acids, and that of tuna by 21 additional amino acids. The degree of similarity reflects the recency of common ancestry. Thus, the inferences from comparative anatomy and other disciplines concerning evolutionary history can be tested in molecular studies of DNA and proteins by examining their sequences of nucleotides and amino acids.

Molecular clues to evolutionary history

The authority of this kind of test is overwhelming; each of the thousands of genes and thousands of proteins contained in an organism provides an independent test of that organism's evolutionary history. Not all possible tests have been performed, but many hundreds have been done, and not one has given evidence contrary to evolution. There is probably no other notion in any field of science that has been as extensively tested and as thoroughly corroborated as the evolutionary origin of living organisms.

The process of evolution

EVOLUTION AS A GENETIC FUNCTION

The concept of natural selection. The central argument of Darwin's theory of evolution starts from the existence of hereditary variation. Experience with animal and plant breeding demonstrates that variations can be developed that are "useful to man." So, reasoned Darwin, variations must occur in nature that are favourable or useful in some way to the organism itself in the struggle for existence. Favourable variations are ones that increase chances for survival and procreation. Those advantageous variations are preserved and multiplied from generation to generation at the expense of less advantageous ones. This is the process known as natural selection. The outcome of the process is an organism that is well adapted to its environment, and evolution often occurs as a consequence.

Natural selection, then, can be defined as the differential reproduction of alternative hereditary variants, determined by the fact that some variants increase the likelihood that the organisms having them will survive and reproduce more successfully than will organisms carrying alternative variants. Selection may be due to differences in survival, in fertility, in rate of development, in mating success, or in any other aspect of the life cycle. All of these differences can be incorporated under the term "differential reproduction" because all result in natural selection to the extent that they affect the number of progeny an organism leaves.

Darwin maintained that competition for limited resources results in the survival of the most effective competitors. But natural selection may occur not only as a result of competition but also as a result of some aspect of the physical environment, such as inclement weather. Moreover,

natural selection would occur even if all the members of a population died at the same age, simply because some of them would have produced more offspring than others.

Darwinian fitness

Natural selection is quantified by a measure called Darwinian fitness, or relative fitness. Fitness in this sense is the relative probability that a hereditary characteristic will be reproduced; that is, the degree of fitness is a measure of the reproductive efficiency of the characteristic.

Biological evolution is the process of change and diversification of living things over time, and it affects all aspects of their lives—morphology, physiology, behaviour, and ecology. Underlying these changes are changes in the hereditary materials. Hence, in genetic terms, evolution consists of changes in the organism's hereditary makeup.

Evolution can be seen as a two-step process. First, hereditary variation takes place; second, selection is made of those genetic variants that will be passed on most effectively to the following generations. Hereditary variation also entails two mechanisms: the spontaneous mutation of one variant to another, and the sexual process that recombines those variants to form a multitude of variations. The variants that arise by mutation or recombination are not transmitted equally from one generation to another. Some may appear more frequently because they are favourable to the organism; the frequency of others may be determined by accidents of chance, called genetic drift.

Genetic variation in populations. *The gene pool.* The gene pool is the sum total of all of the genes and combinations of genes that occur in a population of organisms of the same species. It can be described by citing the frequencies of the alternative genetic constitutions. Consider, for example, a particular gene (which geneticists call a locus), such as the one determining the MN blood groups in humans. One form of the gene codes for the M blood group, while the other form codes for the N blood group; different forms of the same gene are called alleles. The gene pool of a particular population is specified by giving the frequencies of the alleles *M* and *N*. Thus, in the United States the *M* allele in Caucasoids occurs with a frequency of 0.539 and the *N* allele with a frequency of 0.461. In other populations, these frequencies are different; the frequency of the *M* allele is 0.917 in Navajo Indians and 0.178 in Australian Aborigines.

The necessity of hereditary variation for evolutionary change to occur can be understood in terms of the gene pool. Assume, for instance, that at the gene locus that codes for the MN blood groups there is no variation; only the *M* allele exists in all individuals. Evolution of the MN blood groups cannot take place in such a population, since the allelic frequencies have no opportunity to change from generation to generation. On the other hand, in populations in which both alleles *M* and *N* are present, evolutionary change is possible.

Genetic variation and rate of evolution. The more genetic variation that exists in a population, the greater the opportunity for evolution to occur. As the number of gene loci that are variable increases and as the number of alleles at each locus becomes greater, the likelihood that some alleles will change in frequency at the expense of their alternates grows. The British geneticist R.A. Fisher mathematically demonstrated a direct correlation between the amount of genetic variation in a population and the rate of evolutionary change by natural selection. This demonstration is embodied in his fundamental theorem of natural selection (1930): "The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."

Experimental vinegar fly populations

This theorem has been confirmed experimentally. One study employed different strains of *Drosophila serrata,* a species of vinegar fly from eastern Australia and New Guinea. Evolution in vinegar flies can be investigated by using "population cages" and finding out how a population changes over many generations. Experimental populations were set up, with the flies living and reproducing in isolated microcosms. Single-strain populations descended from flies collected either in Popondetta, New Guinea, or in Sydney, Australia; and a mixed population was established by crossing these two strains of flies. The mixed population had the greater initial genetic variation, since it was started by combining two different single-strain populations. To encourage rapid evolutionary change, the populations were manipulated in such a way that there was intense competition among the flies for food and space. Adaptation to the experimental environment was measured by periodically counting the number of individuals in the populations.

Two results deserve notice. First, the mixed population had, at the end of the experiment, more flies than the single-strain populations. Second, and more relevant, the number of flies increased at a faster rate in the mixed population than in the single-strain populations. Evolutionary adaptation to the environment occurred in both types of population; both were able to maintain higher numbers as the generations progressed. But the rate of evolution was more rapid in the mixed group than in the single-strain groups. The greater initial amount of genetic variation made possible a faster rate of evolution.

Measuring gene variability. Because a population's potential for evolving is determined by its genetic variation, evolutionists are interested in discovering the extent of such variation in natural populations. It is readily apparent that plant and animal species are heterogeneous in all sorts of ways; in the flower colours and growth habits of plants, for instance, or the shell shapes and banding patterns of snails. Differences are more readily noticed among humans—in facial features, hair and skin colour, height and weight—but morphological differences are present in all groups of organisms. One problem with morphological variation is that it is not known how much is due to genetic factors and how much may result from environmental influences.

Animal and plant breeders select for their experiments individuals or seeds that excel in desired attributes—in the protein content of corn, for example, or the milk yield of cows. The selection is repeated generation after generation. If the population changes in the direction favoured by the breeder, it becomes clear that the original stock possessed genetic variation with respect to the selected trait. The results of artificial selection are impressive. Selection for high oil content in corn (maize) increased the oil content from less than 5 percent to more than 19 percent in 76 generations, while selection for low oil content reduced it to below 1 percent. Thirty years of selection for increased egg production in a flock of White Leghorn chickens increased the average yearly output of a hen from 125.6 to 249.6 eggs. Artificial selection has produced an endless variety of dog, cat, and horse breeds. The plants that humans grow for food and fibre and the animals they breed for food and transportation are all products of age-old or modern-day artificial selection.

Artificial selection

The success of artificial selection for virtually every trait and every organism in which it has been tried suggests that genetic variation is pervasive throughout natural populations. But evolutionists like to go one step further and obtain quantitative estimates. Only since the 1960s, with the advances of molecular biology, have geneticists developed methods for measuring the extent of genetic variation in populations or among species of organisms. These methods consist essentially of taking a sample of genes and finding out how many are variable and how variable each one is. One simple way of measuring the variability of a gene locus is to ascertain what proportion of the individuals in a population are "heterozygotes" at that locus. In a heterozygous individual, the two genes for a trait, one received from the mother and the other from the father, are different. The proportion of heterozygotes in the population is, therefore, the same as the probability that two genes taken at random from the gene pool are different.

Techniques for determining heterozygosity have been used to investigate numerous species of plants and animals. Typically, insects and other invertebrates are more varied genetically than mammals and other vertebrates; and plants bred by outcrossing exhibit more variation than those bred by self-pollination. But the amount of genetic variation is in any case astounding. Consider as an example humans, whose level of variation is about the same as that of other mammals. The human heterozy-

gosity value is stated as $H = 0.067$, which means that an individual is heterozygous at 6.7 percent of his genes. It is not known how many gene loci there are in humans, but estimates range from 30,000 to 100,000. Assuming the lower estimate, a person would be heterozygous at $30{,}000 \times 0.067 = 2{,}010$ gene loci. An individual heterozygous at one locus (*Aa*) can produce two different kinds of sex cells, or gametes, one with each allele (*A* and *a*); an individual heterozygous at two loci (*AaBb*) can produce four kinds of gametes (*AB, Ab, aB,* and *ab*); an individual heterozygous at *n* loci can potentially produce 2^n different gametes. Therefore, a typical human individual has the potential to produce $2^{2{,}010}$, or approximately 10^{605} (1 with 605 zeros following), different kinds of gametes. But that number is much larger than the estimated number of atoms in the universe, 10^{76}, which is trivial by comparison.

Genetic uniqueness of individuals

It is clear, then, that every sex cell produced by a human being is genetically different from every other sex cell and, therefore, that no two persons who ever existed or will ever exist are likely to be genetically identical—with the exception of identical twins, which develop from a single fertilized ovum. The same conclusion applies to all organisms that reproduce sexually; every individual represents a unique genetic configuration that will never be repeated again. This enormous reservoir of genetic variation in natural populations provides virtually unlimited opportunities for evolutionary change in response to the environmental constraints and the needs of the organisms.

The origin of genetic variation: mutations. Life originated about 3,500,000,000 years ago in the form of primordial organisms that were very simple and very small. All living things have evolved from these lowly beginnings. At present there are more than 2,000,000 known species, which are widely diverse in size, shape, and way of life, as well as in the DNA sequences that contain their genetic information. What has produced the pervasive genetic variation within natural populations and the genetic differences among species? There must be some evolutionary means by which existing DNA sequences are changed and new sequences are incorporated into the gene pools of species.

The information encoded in the nucleotide sequence of DNA is, as a rule, faithfully reproduced during replication, so that each replication results in two DNA molecules that are identical to each other and to the parent molecule. But heredity is not a perfectly conservative process; otherwise, evolution could not have taken place. Occasionally "mistakes," or mutations, occur in the DNA molecule during replication, so that daughter cells differ from the parent cells in the sequence or in the amount of DNA. A mutation first appears on a single cell of an organism, but it is passed on to all cells descended from the first. Mutations can be classified into two categories: gene, or point, mutations, which affect only a few nucleotides within a gene; and chromosomal mutations, which either change the number of chromosomes or change the number or arrangement of genes on a chromosome.

Gene mutations. A gene mutation occurs when the nucleotide sequence of the DNA is altered and a new sequence is passed on to the offspring. The change may be either a substitution of one or a few nucleotides for others or an insertion or deletion of one or a few pairs of nucleotides.

The four nucleotide bases of DNA are represented by the letters A, C, G, and T. A gene that bears the code for constructing a protein molecule consists of a sequence of several thousand nucleotides, so that each segment of three nucleotides—called a codon—codes for one particular amino acid in the protein. The nucleotide sequence in the DNA is first transcribed onto a molecule of messenger RNA (ribonucleic acid). The RNA, with a slightly different code (represented by the letters A, C, G, and U), bears the message that determines which amino acid will be inserted into the protein's chain. Substitutions in the nucleotide sequence of a structural gene may result in changes in the amino acid sequence of the protein, although this is not always the case. The genetic code is redundant in that different triplets may hold the code for the same amino acid. Consider the triplet UUA in the messenger RNA, which codes for the amino acid leucine. If the first U is replaced by C, the triplet will still code for leucine; but if it is replaced by G, it will code for valine instead.

Changing the genetic instruction

A nucleotide substitution in the DNA that results in an amino acid substitution in the corresponding protein may or may not severely affect the biological function of the protein. Some nucleotide substitutions change a codon for an amino acid into a stop signal, and those mutations are likely to have harmful effects. If, for instance, the second U in the UUA triplet is replaced by A, the triplet becomes UAA, a "terminator" codon; the result is that the following triplets in the DNA sequence are not translated into amino acids.

Additions or deletions of nucleotide pairs within the DNA sequence of a structural gene often result in a greatly altered sequence of amino acids in the coded protein. The addition or deletion of one or two nucleotide pairs shifts the "reading frame" of the nucleotide sequence all along the way from the point of the insertion or deletion to the end of the molecule. To illustrate, assume that a DNA segment is read as . . . CAT-CAT-CAT-CAT-CAT. . . . If a nucleotide base, say T, is inserted after the C of the first triplet, the segment will then be read as . . . CTA-TCA-TCA-TCA-TCA. . . . From the point of the insertion onward, the sequence of encoded amino acids is altered. If, however, a total of three nucleotide pairs is either added or deleted, the original reading frame will be restored in the rest of the sequence. Additions or deletions of nucleotide pairs in numbers other than three or multiples of three are called frameshift mutations.

Gene mutations can occur spontaneously; that is, without being intentionally caused by humans. They can also be induced by ultraviolet light, X rays, and other high-frequency radiations, as well as by exposure to certain mutagenic chemicals, such as mustard gas. The consequences of gene mutations may range from negligible to lethal. Mutations that change one or even several amino acids may have a small or undetectable effect on the organism's ability to survive and reproduce if the essential biological function of the coded protein is not hindered. But where an amino acid substitution affects the active site of an enzyme or modifies in some other way an essential function of a protein, the impact may be severe.

Newly arisen mutations are more likely to be harmful than beneficial to their carriers, because mutations are random events with respect to adaptation; that is, their occurrence is independent of any possible consequences. The allelic variants present in an existing population have already been subject to natural selection. They are present in the population because they improve the adaptation of their carriers, and their alternative alleles have been eliminated or kept at low frequencies by natural selection. A newly arisen mutant is likely to have been preceded by an identical mutation in the previous history of a population; if the previous mutant no longer exists in the population, that will be a sign that the new mutant is not beneficial to the organism and is likely to be eliminated also.

This proposition can be illustrated with an analogy. Consider an English sentence, whose words have been chosen because together they express a certain idea. If single letters or words are replaced with others at random, most changes will be unlikely to improve the meaning of the sentence; very likely they will destroy it. The nucleotide sequence of a gene has been "edited" into its present form by natural selection because it "makes sense." If the sequence is changed at random, the "meaning" rarely will be improved and often will be hampered or destroyed.

Occasionally, however, a new mutation may increase the organism's adaptation. The probability of such an event's happening is greater when organisms colonize a new territory or when environmental changes confront a population with new challenges. In these cases, the established adaptation of a population is less than optimal, and there is greater opportunity for new mutations to be better adaptive. The consequences of mutations depend on the environment. Increased melanin pigmentation may be advantageous to inhabitants of tropical Africa, where dark skin protects them from the Sun's ultraviolet radiation; but it is not beneficial in Scandinavia, where the intensity

Adaptation to environmental challenges

of sunlight is low and light skin facilitates the synthesis of vitamin D.

Mutation rates have been measured in a great variety of organisms, mostly for mutants that exhibit conspicuous effects. Mutation rates are generally lower in bacteria and other microorganisms than in more complex species. In humans and other multicellular organisms, the rate typically ranges from about one per 100,000 to one per 1,000,000 gametes. There is, however, considerable variation from gene to gene as well as from organism to organism.

Although mutation rates are low, new mutants appear continuously in nature, because there are many individuals in every species and many gene loci in every individual. The process of mutation provides each generation with many new genetic variations. Thus, it is not surprising to see that when new environmental challenges arise, species are able to adapt to them. More than 200 insect and rodent species, for example, have developed resistance to the pesticide DDT in different parts of the world where spraying has been intense. Although the insects had never before encountered this synthetic compound, they adapted to it rapidly by means of mutations that allowed them to survive in its presence. Similarly, many species of moths and butterflies in industrialized regions have shown an increase in the frequency of individuals with dark wings in response to environmental pollution, an adaptation known as industrial melanism.

Chromosomal mutations. The chromosomes, which carry the hereditary material, or DNA, are contained in the nucleus of each cell. Chromosomes come in pairs, with one member of each pair inherited from each parent. The two members of a pair are called homologous chromosomes. Each cell of an organism and all individuals of the same species have, as a rule, the same number of chromosomes. The reproductive cells (gametes) are an exception; they have only half as many chromosomes as the body (somatic) cells. But the number, size, and organization of chromosomes varies among species. The parasitic nematode *Parascaris univalens* has only one pair of chromosomes, whereas many species of butterflies have more than 100 pairs and some ferns more than 600. Even closely related organisms may vary considerably in the number of chromosomes; species of spiny rats of the South American genus *Proechimys* range from 12 to 31 chromosome pairs. Changes in the number, size, or organization of chromosomes are termed chromosomal mutations, chromosomal abnormalities, or chromosomal aberrations.

Changes in the number and structure of chromosomes

Changes in the number of chromosomes may occur by fusion of two chromosomes into one, by fission of one chromosome into two, or by addition or subtraction of one or more whole chromosomes or sets of chromosomes (polyploidy). Changes in the structure of chromosomes may occur by inversion, when a chromosomal segment rotates 180° within the same location; by duplication, when a segment is added; by deletion, when a segment is lost; or by translocation, when a segment changes from one location to another in the same or a different chromosome. These are the processes by which chromosomes evolve. Inversions, translocations, fusions, and fissions do not change the amount of DNA. The importance of these mutations in evolution is that they change the linkage relationships between genes. Genes that were closely linked to each other become separated and vice versa.

DYNAMICS OF GENETIC CHANGE

Genetic equilibrium: the Hardy–Weinberg law. Genetic variation is present throughout natural populations of organisms. This variation is sorted out in new ways in each generation by the process of sexual reproduction, which recombines the chromosomes inherited from the two parents during the formation of the gametes that produce the following generation. But heredity by itself does not change gene frequencies. This principle is stated by the Hardy–Weinberg law, so called because it was independently discovered in 1908 by the English mathematician G.H. Hardy and the German physician Wilhelm Weinberg.

The Hardy–Weinberg law describes the genetic equilibrium in a population by means of an algebraic equation. It states that genotypes (the genetic constitution of individual organisms) exist in certain frequencies that are a simple function of the allelic frequencies; namely, the square expansion of the sum of the allelic frequencies.

If there are two alleles, A and a, at a gene locus, three genotypes will be possible, AA, Aa, and aa. If the frequencies of the two alleles are p and q, respectively, the equilibrium frequencies of the three genotypes will be given by $(p+q)^2 = p^2 + 2pq + q^2$, for AA, Aa, and aa, respectively. The genotype equilibrium frequencies for any number of alleles are derived in the same way. If there are three alleles, A_1, A_2, and A_3, with frequencies p, q, and r, the equilibrium frequencies corresponding to the six possible genotypes (shown in parentheses) will be calculated as follows:

$$(p+q+r)^2 = p^2(A_1A_1) + q^2(A_2A_2) + r^2(A_3A_3) + 2pq(A_1A_2) + 2pr(A_1A_3) + 2qr(A_2A_3).$$

Table 1 shows how the law operates in a situation with just two alleles. At the top and to the left are the frequencies in the parental generation of the two alleles, p for A and q for a. As shown at the lower right, the probabilities of the three possible genotypes in the following generation are products of the probabilities of the corresponding alleles in the parents. The probability of genotype AA among the progeny is the probability p that allele A will be present in the paternal gamete multiplied by the probability p that allele A will be present in the maternal gamete, or p^2. Similarly, the probability of the genotype aa is q^2. The genotype Aa can arise by getting A from the father and a from the mother, which will occur with a frequency pq, or by getting a from the father and A from the mother, which also has a probability of pq; this gives a total probability of $2pq$ for the frequency of the Aa genotype in the progeny.

Genotype probabilities

Table 1: The Hardy–Weinberg Law Applied to Two Alleles

paternal gametic frequencies	maternal gametic frequencies	
	$p(A)$	$q(a)$
$p(A)$	$p^2(AA)$	$pq(Aa)$
$q(a)$	$pq(Aa)$	$q^2(aa)$

There is no change in the allele equilibrium frequencies from one generation to the next. The frequency of the A allele among the offspring is the frequency of the AA genotype (because all alleles in these individuals are A alleles) plus half the frequency of the Aa genotype (because half the alleles in these individuals are A alleles), or $p^2 + pq = p(p+q) = p$ (because $p+q=1$). Similarly, the frequency of the a allele among the offspring is given by $q^2 + pq = q(q+p) = q$. These are precisely the frequencies of the alleles in the parents.

The genotype equilibrium frequencies are obtained by the Hardy–Weinberg law on the assumption that there is random mating; that is, the probability of a particular kind of mating is the same as the frequency of the genotypes of the two mating individuals. For example, the probability of an AA female mating with an aa male must be p^2 (the frequency of AA) times q^2 (the frequency of aa). Random mating can occur with respect to most gene loci even though mates may be chosen according to particular characteristics. People, for example, choose their spouses according to all sorts of preferences concerning looks, personality, and the like. But concerning the majority of genes, people's marriages are essentially random.

Assortative, or selective, mating takes place when the choice of mates is not random. Marriages in the United States, for example, are assortative with respect to racial features, so that Negroes, Asians, and Caucasoids marry members of their own racial group more often, and people from a different racial group less often, than would be expected from random mating. Consider a community in which 80 percent of the population are white and 20 percent are black. With random mating, 32 percent ($2 \times 0.80 \times 0.20 = 0.32$) of all marriages would be interracial, whereas only 4 percent ($0.20 \times 0.20 = 0.04$) would be marriages between two blacks. These expectations depart from typical observations. The most extreme form of as-

sortative mating is self-fertilization, which occurs rarely in animals but is a common form of reproduction in many plant groups.

Assumptions of the Hardy–Weinberg law

The Hardy–Weinberg law assumes that gene frequencies remain constant from generation to generation—that there is no gene mutation or natural selection and that populations are very large. But these assumptions are not correct; indeed, if they were, evolution could not occur. Why, then, is the Hardy–Weinberg law significant if its assumptions do not hold true in nature? The answer is that the Hardy–Weinberg law plays in evolutionary studies a role similar to that of Newton's first law of motion in mechanics. Newton's first law says that a body not acted upon by a net external force remains at rest or maintains a constant velocity. In fact, there are always external forces acting upon physical objects, but the first law provides the starting point for the application of other laws. Similarly, organisms are subject to mutation, selection, and other processes that change gene frequencies, but the effects of these processes can be calculated by using the Hardy–Weinberg law as the starting point.

Processes of gene frequency change. *Mutation.* The allelic variations that make evolution possible are generated by the process of mutation; but new mutations change gene frequencies very slowly, since mutation rates are low. Assume that the gene allele A_1 mutates to allele A_2 at a rate m per generation, and that at a given time the frequency of A_1 is p. In the next generation, a fraction m of all A_1 alleles become A_2 alleles. The frequency of A_1 in the next generation will then be reduced by the fraction of mutated alleles (pm), or $p_1 = p - pm = p(1 - m)$. After t generations, the frequency of A_1 will be $p_t = p(1 - m)^t$.

If the mutations continue, the frequency of A_1 alleles will gradually decrease, because a fraction of them change every generation to A_2. If the process continues indefinitely, the A_1 allele will eventually disappear, although the process is slow. If the mutation rate is 10^{-5} (1 in 100,000) per gene per generation, about 2,000 generations will be required to change the frequency of A_1 from 0.50 to 0.49 and about 10,000 generations to change it from 0.10 to 0.09.

Moreover, gene mutations are reversible: the allele A_2 may also mutate to A_1. Assume that A_1 mutates to A_2 at a rate m, as before, and that A_2 mutates to A_1 at a rate n per generation. If at a certain time the frequencies of A_1 and A_2 are p and q, respectively, after one generation the frequency of A_1 will be $p_1 = p - pm + qn$. A fraction pm of allele A_1 changes to A_2, but a fraction qn of the A_2 alleles changes to A_1. The conditions for equilibrium occur when $pm = qn$, or $p = m/(m + n)$. Suppose that the mutation rates are $m = 10^{-6}$ and $n = 10^{-5}$; then, at equilibrium, $p = 10^{-6}/(10^{-6} + 10^{-5}) = 1/(1 + 10) = 0.09$, and $q = 0.91$.

Changes in gene frequencies due to mutation occur, therefore, at even slower rates than was suggested above, because forward and backward mutations counteract each other. In any case, allelic frequencies usually are not in mutational equilibrium, because some alleles are favoured over others by natural selection. The equilibrium frequencies are then decided by the interaction between mutation and selection, with selection usually having the greater consequence.

Migration. Gene flow, or gene migration, takes place when individuals migrate from one population to another and interbreed with its members. Gene frequencies are not changed for the species as a whole, but they change locally whenever different populations have different allele frequencies. In general, the greater the difference in allele frequencies between the resident and the migrant individuals, and the larger the number of migrants, the greater effect the migrants have in changing the genetic constitution of the resident population.

Effect of migrants on a resident population

Suppose that a proportion of all reproducing individuals in a population are migrants and that the frequency of allele A_1 is p in the population but p_m among the migrants. The change in gene frequency, Δp, in the next generation will be $\Delta p = m(p_m - p)$. If the migration rate persists for a number t of generations, the frequency of A_1 will be given by $p_t = (1 - m)^t(p - p_m) + p_m$.

Genetic drift. Gene frequencies can change from one generation to another by a process of pure chance known as genetic drift. This occurs because populations are finite in numbers, and thus the frequency of a gene may change in the following generation by accidents of sampling, just as it is possible to get more or less than 50 "heads" in 100 throws of a coin simply by chance.

The magnitude of the gene frequency changes due to genetic drift is inversely related to the size of the population; the larger the number of reproducing individuals, the smaller the effects of genetic drift. This inverse relationship between sample size and magnitude of sampling errors can be illustrated by referring again to tossing a coin. When a penny is tossed twice, two heads are not surprising. But it will be surprising, and suspicious, if 20 tosses all yield heads. The proportion of heads obtained in a series of throws approaches closer to 0.5 as the number of throws grows larger.

The relationship is the same in populations, although the important value here is not the actual number of individuals in the population but the "effective" population size. This means the number of individuals that produce offspring, because only reproducing individuals transmit their genes to the following generation. It is not unusual, in plants as well as animals, for some individuals to have large numbers of progeny while others have none. In marine seals, antelopes, baboons, and many other mammals, for example, a dominant male may keep a large harem of females at the expense of many other males who can find no mates. It often happens that the effective population size is substantially smaller than the number of individuals in any one generation.

The effects of genetic drift in changing gene frequencies from one generation to the next are quite small in most natural populations, which generally consist of thousands of reproducing individuals. The effects over many generations are more important. Indeed, in the absence of other processes of change (such as natural selection and mutation), populations would eventually become fixed, having one allele at each locus after the gradual elimination of all others. With genetic drift as the only force in operation, the probability of a given allele eventually reaching a frequency of 1 would be precisely the frequency of the allele; that is, an allele with a frequency of 0.8 would have an 80 percent chance of ultimately becoming the only allele present in the population. The process would, however, take a long time, because increases and decreases are likely to alternate with equal probability. More important, natural selection and other processes change gene frequencies in deterministic ways, so that no allele has an opportunity to become fixed as a consequence of genetic drift alone.

Genetic drift can have important evolutionary consequences when a new population becomes established by only a few individuals—a phenomenon known as the founder principle. Islands, lakes, and other isolated ecological sites are often colonized by one or very few seeds or animals of a species, which are transported there passively by wind, in the fur of larger animals, or in some other way. The allelic frequencies present in these few colonizers are likely to differ at many loci from those in the population they came from, thus having a lasting impact on the evolution of the new population. The founder principle is one reason that species in neighbouring islands, such as those in the Hawaiian archipelago, are often more heterogeneous than species in comparable continental areas adjacent to one another.

The founder principle

Climatic or other conditions, if unfavourable, may on occasion drastically reduce the number of individuals in a population and even threaten it with extinction. Such occasional reductions are called population bottlenecks. The populations may later recover their typical size, but the allelic frequencies may have been considerably altered, thereby affecting the future evolution of the species. Bottlenecks are more likely in relatively large animals and plants than in smaller ones, because populations of large organisms typically consist of fewer individuals. Primitive human populations of the past were subdivided into many small tribes that were time and again decimated by disease, war, and other disasters. Differences among current human populations in the allele frequencies of many genes—such as those determining the ABO and

other blood groups—may have arisen at least in part as a consequence of bottlenecks in ancestral populations. Persistent population bottlenecks may reduce the overall genetic variation so greatly as to alter future evolution and endanger the survival of the species. A well-authenticated case is that of the cheetah, where no allelic variation whatsoever has been found among the many scores of gene loci studied.

THE OPERATION OF NATURAL SELECTION IN POPULATIONS

Natural selection as a process of genetic change. Natural selection refers to any reproductive bias favouring some genes or genotypes over others. Natural selection promotes the adaptation of organisms to the environments in which they live; any hereditary variant that improves the ability to survive and reproduce in an environment will increase in frequency over the generations, precisely because the organisms carrying such a variant will leave more descendants than those lacking it. Hereditary variants, favourable or not to the organisms, arise by mutation. Unfavourable ones are eventually eliminated by natural selection; their carriers leave no descendants or leave fewer than those carrying alternative variants. Favourable mutations accumulate over the generations. The process continues indefinitely because the environments that organisms live in are forever changing. Environments change physically—in their climate, physical configuration, and so on—but also biologically, because the predators, parasites, and competitors with which an organism interacts are themselves evolving.

The ever-changing environment

Mutation, migration, and drift are random processes with respect to adaptation; they change gene frequencies without regard for the consequences that such changes may have in the ability of the organisms to survive and reproduce. If these were the only processes of evolutionary change, the organization of living things would gradually disintegrate. The effects of such processes alone would be analogous to those of a mechanic who changed parts in a motorcar engine at random, with no regard for the role of the parts in the engine. Natural selection keeps the disorganizing effects of mutation and other processes in check because it multiplies beneficial mutations and eliminates harmful ones.

Natural selection accounts not only for the preservation and improvement of the organization of living beings but also for their diversity. In different localities or in different circumstances, natural selection favours different traits, precisely those that make the organisms well adapted to their particular circumstances and ways of life.

The parameter used to measure the effects of natural selection is fitness, which can be expressed as an absolute or as a relative value. Consider a population consisting at a certain locus of three genotypes: A_1A_1, A_1A_2, and A_2A_2. Assume that on the average each A_1A_1 and each A_1A_2 individual produces one offspring, but that each A_2A_2 individual produces two. One could use the average number of progeny left by each genotype as a measure of that genotype's absolute fitness and calculate the changes in gene frequency that would occur over the generations (this, of course, requires knowing how many of the progeny survive to adulthood and reproduce). Evolutionists, however, find it mathematically more convenient to use relative fitness values—which they represent with the letter w—in most calculations. They usually assign the value 1 to the genotype with the highest reproductive efficiency and calculate the other relative fitness values proportionally. For the example just used, the relative fitness of the A_2A_2 genotype would be $w = 1$ and that of each of the other two genotypes would be $w = 0.5$. A parameter related to fitness is the selection coefficient, often represented with the letter s, which is defined as $s = 1 - w$. The selection coefficient is a measure of the reduction in fitness of a genotype. The selection coefficients in the example are $s = 0$ for A_2A_2 and $s = 0.5$ for A_1A_1 and A_1A_2.

The different ways in which natural selection affects gene frequencies are illustrated by the following examples.

Selection against one of the homozygotes. Suppose that one homozygous genotype, say A_2A_2, has lower fitness than the other two genotypes, A_1A_1 and A_1A_2. (This is the situation in many human diseases, such as phenylketonuria (PKU) and sickle-cell anemia. The heterozygotes and the homozygotes for the normal allele have equal fitness, higher than that of the homozygotes for the deleterious allele.) Call the fitness of these homozygotes $1 - s$ (the fitness of the other two genotypes is 1), and let p be the frequency of the normal allele (A_1) and q the frequency of the deleterious allele (A_2). It can be shown that the frequency of A_2 will decrease each generation by an amount given by $\Delta q = -spq^2/(1 - sq^2)$. The deleterious allele, A_2, will continuously decrease in frequency until it is eliminated. The rate of elimination is fastest when $s = 1$ (*i.e.*, $w = 0$); this occurs with fatal diseases, such as untreated PKU, when the homozygotes die before the age of reproduction.

Because of new mutations, the elimination of a deleterious allele is never complete. A dynamic equilibrium frequency will exist when the number of new alleles produced by mutation is the same as the number eliminated by selection. If the mutation rate at which the deleterious allele arises is u, the equilibrium frequency for a deleterious allele that is recessive is given approximately by $q = \sqrt{u/s}$, which, if $s = 1$, reduces to $q = \sqrt{u}$.

The mutation rate for many human recessive diseases is about 1 in 100,000 ($u = 10^{-5}$). If the disease is fatal, the equilibrium frequency becomes $q \cong \sqrt{10^{-5}} = 0.003$, or about 1 recessive lethal allele for every 300 normal alleles. That is roughly the frequency in human populations of alleles that in homozygous individuals, like those with PKU, cause death before adulthood. The equilibrium frequency for a deleterious, but not lethal, recessive allele is much higher. Albinism, for example, is due to a recessive gene. The reproductive efficiency of albinos is, on average, about 0.9 that of normal individuals. Therefore, $s = 0.1$ and $q = \sqrt{u/s} = \sqrt{10^{-5}/10^{-1}} = 0.01$, or 1 in 100 genes rather than 1 in 300 as for a lethal allele.

For deleterious dominant alleles, the mutation-selection equilibrium frequency is given by $p = u/s$, which for fatal genes becomes $p = u$. If the gene is lethal even in single copy, all the genes are eliminated by selection in the same generation in which they arise, and the frequency of the gene in the population is the frequency with which it arises by mutation. One deleterious condition that is caused by a dominant allele present at low frequencies in human populations is achondroplasia. Because of abnormal growth of the long bones, achondroplastics have short, squat, often deformed limbs, along with bulging skulls. The mutation rate from the normal allele to the achondroplasia allele is about 5×10^{-5}. Achondroplastics reproduce only 20 percent as efficiently as normal individuals; hence, $s = 0.8$. The equilibrium frequency of the allele can therefore be calculated as $p = u/s = 6.25 \times 10^{-5}$.

Overdominance. In many instances the heterozygotes have a higher degree of fitness than the homozygotes for one or the other allele. This situation, known as heterosis or overdominance, leads to the stable coexistence of both alleles in the population and, hence, contributes to the widespread genetic variation found in populations of most organisms. The model situation is:

Greater fitness of heterozygotes

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness (w)	$1 - s$	1	$1 - t$

It is assumed that s and t are positive numbers between 0 and 1, so that the fitnesses of the two homozygotes are somewhat less than 1. It is not difficult to show that the change in frequency per generation of allele A_2 is $\Delta q = pq(sp - tq)/(1 - sp^2 - tq^2)$. An equilibrium will exist when $\Delta q = 0$ (gene frequencies no longer change); this will happen when $sp = tq$, at which the numerator of the expression for Δq will be 0. The condition $sp = tq$ can be rewritten as $s(1 - q) = tq$ (when $p + q = 1$), which leads to $q = s/(s + t)$. If the fitnesses of the two homozygotes are known, it is possible to infer the allele equilibrium frequencies.

A colour polymorphism in the marine copepod *Tisbe reticulata* is one of many well-investigated examples of overdominance in animals. Three colour morphs found in the lagoon of Venice are known as violacea (homozygous

genotype $V^V V^V$), maculata (homozygous genotype $V^M V^M$), and violacea-maculata (heterozygous genotype $V^V V^M$). The colour polymorphism persists in the lagoon because the heterozygotes survive better than either one of the two homozygotes. In laboratory experiments, the fitness of the three genotypes depends on the degree of crowding, as shown by the following comparison of their relative fitnesses:

Genotype	Fitness in low crowding	Fitness in high crowding
$V^V V^V$	0.89	0.66
$V^V V^M$	1	1
$V^M V^M$	0.90	0.62

The greater the crowding—with more competition for resources—the greater the superiority of the heterozygotes.

Sickle-cell anemia

A particularly interesting example of heterozygote superiority among humans is provided by the gene responsible for sickle-cell anemia. Human hemoglobin in adults is for the most part hemoglobin A, a four-component molecule consisting of two α and two β hemoglobin chains. The normal β hemoglobin chain consists of 146 amino acids and is coded for by the gene Hb^A. A mutant allele of this gene, Hb^S, causes the β chain to have in the sixth position the amino acid valine instead of glutamic acid. This seemingly minor substitution modifies the properties of hemoglobin so that homozygotes with the mutant allele, $Hb^S Hb^S$, suffer from a severe form of anemia that in most cases leads to death before the age of reproduction.

The high frequency of the Hb^S allele in some African and Asian populations formerly was puzzling because the severity of the anemia, representing a strong natural selection against homozygotes, should have eliminated the defective allele. The situation was understood after it was noticed that the Hb^S allele occurred at high frequency precisely in regions of the world where a particularly severe form of malaria (caused by the parasite *Plasmodium falciparum*) was endemic. It was hypothesized that the heterozygotes, $Hb^A Hb^S$, were resistant to malaria, whereas the homozygotes $Hb^A Hb^A$ were not. In malaria-infested regions, then, the heterozygotes survived better than either one of the two homozygotes, which were more likely to die from either malaria ($Hb^A Hb^A$ homozygotes) or anemia ($Hb^S Hb^S$ homozygotes). This hypothesis has been confirmed in various ways. Most significant is that most hospital patients suffering from severe or fatal forms of malaria are homozygotes $Hb^A Hb^A$. In a study of 100 children who died from malaria, only one was found to be a heterozygote, while 22 were expected to be so according to the frequency of the Hb^S allele in the population.

Table 2 shows how the relative fitness of the three β-chain genotypes can be calculated from their distribution among the Yoruba people of Ibadan, Nigeria. The frequency of the Hb^S allele among adults is estimated as $q = 0.1232$. According to the Hardy–Weinberg law, the three genotypes will be formed at conception in the frequencies p^2, $2pq$, and q^2, which are the expected frequencies given in Table 2. The ratios of the observed frequencies among adults to the expected frequencies give the relative survival efficiency of the three genotypes. These are divided by their largest value (1.12) in order to obtain the relative fitness of the genotypes. Sickle-cell anemia reduces the probability of survival of the $Hb^S Hb^S$ homozygotes to 13 percent of that of the heterozygotes; but the homozygotes for the normal allele, $Hb^A Hb^A$, have their survival probability reduced by malaria infection to 88 percent.

Frequency-dependent selection. The fitness of genotypes can change when the environmental conditions change. White fur may be protective to a bear living on the Arctic snows, but not to one living in a California forest; there an allele coding for brown pigmentation may be favoured over one that codes for white. The environment of an organism includes not only the climate and other physical features, but also the organisms of the same or different species with which it is associated.

Effect of population density on fitness

Changes in genotypic fitness are associated with the density of the organisms present. Insects and other short-lived organisms experience enormous yearly oscillations in density. Some genotypes may possess high fitness in the spring, when the population is rapidly expanding, because such genotypes yield more prolific individuals. Other genotypes may be favoured during the summer, when populations are dense, because these genotypes make for better competitors, more successful at securing limited food resources. Still others may be at an advantage during the long winter months, because they increase the population's hardiness, or ability to withstand the inclement conditions that kill most members of the other genotypes.

The fitness of genotypes can also vary according to their relative numbers, and genotype frequencies may change as a consequence. This is known as frequency-dependent selection. Particularly interesting is the situation in which genotypic fitnesses are inversely related to their frequencies. Assume that two genotypes, *A* and *B*, have fitnesses related to their frequencies in such a way that the fitness of either genotype increases when its frequency decreases and vice versa. When *A* is rare, its fitness is high, and therefore *A* increases in frequency; but as it becomes more and more common, the fitness of *A* gradually decreases, so that its increase in frequency eventually comes to a halt. A stable polymorphism occurs at the frequency where the two genotypes, *A* and *B*, have identical fitnesses.

In natural populations of animals and plants frequency-dependent selection is very common and may contribute importantly to the maintenance of genetic polymorphism. In the vinegar fly *Drosophila pseudoobscura*, for example, three genotypes exist at the gene locus coding for the enzyme malate dehydrogenase: the homozygous *SS* and *FF* and the heterozygous *SF*. When the *SS* homozygotes represent 90 percent of the population, they have a fitness about two-thirds that of the heterozygotes, *SF;* but when the *SS* homozygotes represent only 10 percent of the population, their fitness is more than double that of the heterozygotes. Similarly, the fitness of the *FF* homozygotes relative to the heterozygotes increases from less than half to nearly double as their frequency goes from 90 to 10 percent. All three genotypes have equal fitnesses when the frequency of the *S* allele, *p*, is about 0.70, so that there is a stable polymorphism with frequencies $p^2 = 0.49$ for *SS*, $2pq = 0.42$ for *SF*, and $q^2 = 0.09$ for *FF*.

Advantage of genotype mixtures

Frequency-dependent selection may arise because the environment is heterogeneous and because different genotypes can better exploit different subenvironments. When a genotype is rare, the subenvironments that it exploits better will be relatively abundant. But as the genotype becomes common, its favoured subenvironment becomes saturated. That genotype must then compete for resources in subenvironments that are optimal for other genotypes. It follows, then, that a mixture of genotypes exploits the environmental resources better than a single genotype. This has been extensively demonstrated. When the three *Drosophila* genotypes mentioned above were mixed in a single population, the average number of individuals that developed per unit of food was 45.6. This was greater than the numbers of individuals that developed when only one of the genotypes was present, which averaged 41.1 for *SS*, 40.2 for *SF*, and 37.1 for *FF*. Plant breeders know that mixed plantings are more productive than single stands, although farmers avoid them for reasons such as increased harvesting costs.

Sexual preferences can also lead to frequency-dependent selection. It has been demonstrated in some insects, birds, mammals, and other organisms that the mates preferred are precisely those that are rare. People also experience this rare-mate advantage; blonds may seem attractively exotic to Latins, or brunets to Scandinavians.

Table 2: Fitnesses of the Three Genotypes at the Sickle-cell Anemia Locus in a Population from Nigeria

	genotype				
	$Hb^A Hb^A$	$Hb^A Hb^S$	$Hb^S Hb^S$	total	frequency of Hb^S
Observed number	9365	2993	29	12,387	
Observed frequency	0.7560	0.2416	0.0023	1	0.1232
Expected frequency	0.7688	0.2160	0.0152	1	0.1232
Survival efficiency	0.98	1.12	0.15		
Relative fitness	0.88	1	0.13		

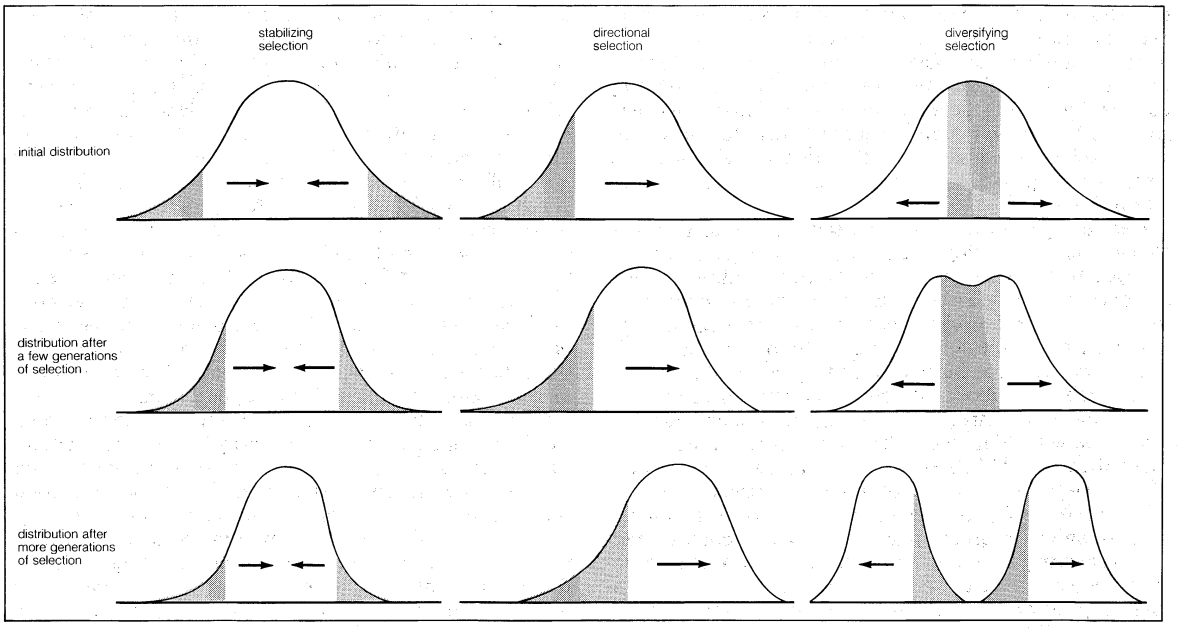


Figure 4: Three types of natural selection showing the effects of each on the distribution of phenotypes within a population. The shaded areas represent the phenotypes against which selection acts. Stabilizing selection acts against phenotypes at both extremes of the distribution, favouring the multiplication of intermediate phenotypes. Directional selection acts against only one extreme of phenotypes, causing a shift in distribution toward the other extreme. Diversifying selection acts against intermediate phenotypes, creating a split in distribution toward each extreme.

Types of selection. *Stabilizing selection.* Natural selection can be studied by analyzing its effects on changing gene frequencies; but it can also be explored by examining its effects on the observable characteristics—or phenotypes—of individuals in a population. Distribution scales of phenotypic traits such as height, weight, number of progeny, or longevity typically show greater numbers of individuals with intermediate values and fewer and fewer toward the extremes (the so-called normal distribution). When individuals with intermediate phenotypes are favoured and extreme phenotypes are selected against, the selection is said to be stabilizing. The range and distribution of phenotypes then remains approximately the same from one generation to another. Stabilizing selection is very common. The individuals that survive and reproduce more successfully are those that have intermediate phenotypic values. Mortality among newborn infants, for example, is highest when they are either very small or very large; infants of intermediate size have a greater chance of surviving.

Stabilizing selection is often noticeable after artificial selection. Breeders choose chickens that produce larger eggs, cows that yield more milk, and corn with higher protein content. But the selection must be continued or reinstated from time to time, even after the desired goals have been achieved. If it is stopped altogether, natural selection gradually takes effect and turns the traits back toward their original intermediate value.

Genetic homeostasis

As a result of stabilizing selection, populations often maintain a steady genetic constitution with respect to many traits. This attribute of populations is called genetic homeostasis.

Directional selection. The distribution of phenotypes in a population sometimes changes systematically in a particular direction. The physical and biological aspects of the environment are continuously changing, and over long periods of time the changes may be substantial. The climate and even the configuration of the land or waters vary incessantly. Changes also take place in the biotic conditions; that is, in the other organisms present, whether predators, prey, parasites, or competitors. Genetic changes occur as a consequence, because the genotypic fitnesses may be shifted so that different sets of alleles are favoured. The opportunity for directional selection also arises when organisms colonize new environments where the conditions are different from those of their original habitat. In addition, the appearance of a new favourable allele or a new genetic combination may prompt directional changes, as the new genetic constitution replaces the preexisting one.

The process of directional selection takes place in spurts. The replacement of one genetic constitution with another changes the genotypic fitnesses at other loci, which then change in their allelic frequencies, thereby stimulating additional changes, and so on in a cascade of consequences.

Directional selection is possible only if there is genetic variation with respect to the phenotypic traits under selection. Natural populations contain large stores of genetic variation, and these are continuously replenished by additional new variants that arise by mutation. The nearly universal success of artificial selection and the rapid response of natural populations to new environmental challenges are evidence that existing variation provides the necessary materials for directional selection.

In modern times, human actions have been an important stimulus to this type of selection. Mankind transforms the environments of many organisms, which rapidly respond to the new environmental challenges through directional selection. Well-known instances are the many cases of insect resistance to pesticides, synthetic substances not present in the natural environment. Whenever a new insecticide is first applied to control a pest, the results are encouraging because a small amount of the insecticide is sufficient to bring the pest organism under control. As time passes, however, the amount required to achieve a certain level of control must be increased again and again until finally it becomes ineffective or economically impractical. This occurs because organisms become resistant to the pesticide through directional selection. The resistance of the housefly, *Musca domestica,* to DDT was first reported in 1947. Resistance to one or more pesticides has now been recorded in more than 100 species of insects.

Industrial melanism

Another example is the phenomenon of industrial melanism—the gradual darkening of the wings of many species of moths and butterflies living in woodlands darkened by industrial pollution. The best investigated case is the peppered moth, *Biston betularia,* of England. Until the middle of the 19th century these moths were uniformly peppered light gray. Darkly pigmented variants were first detected in 1848 in Manchester, Eng., and shortly afterward in other industrial regions where the vegetation was blackened by soot and other pollutants. By the middle of the 20th century the dark varieties had almost completely

replaced the lightly pigmented forms in many polluted areas, while in unpolluted regions light moths continued to be the most common. The shift from light to dark moths was an example of directional selection brought about by bird predators. On lichen-covered tree trunks, the light-gray moths are well camouflaged, whereas the dark ones are conspicuously visible and therefore fall victim to the birds. The opposite is the case on trees darkened by pollution (Figure 5).

Over geologic time, directional selection leads to major changes in morphology and ways of life. Evolutionary changes that persist in a more or less continuous fashion over long periods of time are known as evolutionary trends. Directional evolutionary changes increased the cranial capacity of the human lineage from the small brain of *Australopithecus,* human ancestors of 3,000,000 years ago, which were about 500 cubic centimetres in volume, to a brain nearly three times as large in modern humans, *Homo sapiens.* The evolution of the horse family from more than 50,000,000 years ago to modern times is another well-studied example of directional selection.

Diversifying selection. Two or more divergent phenotypes in an environment may be favoured simultaneously by diversifying selection. No natural environment is homogeneous; rather, the environment of any plant or animal population is a mosaic consisting of more or less dissimilar subenvironments. There is heterogeneity with respect to climate, food resources, and living space. Also, the heterogeneity may be temporal, with change occurring over time, as well as spatial, with dissimilarity found in different areas. Species cope with environmental heterogeneity in diverse ways. One strategy is the selection of a generalist genotype that is well adapted to all of the subenvironments encountered by the species. Another strategy is genetic polymorphism, the selection of a diversified gene pool that yields different genotypes, each adapted to a specific subenvironment.

There is no single plan that prevails in nature. Sometimes the most efficient strategy is genetic monomorphism to confront temporal heterogeneity but polymorphism to confront spatial heterogeneity. If the environment changes in time, if it is unstable relative to the life span of the organisms, each individual will have to face diverse environments appearing one after the other. A series of genotypes, each well adapted to one or another of the conditions that prevail at various times, will not succeed very well, because each organism will fare well at one period of its life but not at others. A better strategy is to have a population with one or a few genotypes that survive well in all the successive environments.

Strategies for adaptation to diverse environments

With respect to spatial heterogeneity, the situation is likely to be different. A single genotype, well adapted to the various environmental patches, is a possible strategy; but a variety of genotypes, with some individuals optimally adapted to each subenvironment, might fare still better. The ability of the population to exploit the environmental patchiness is thereby increased. Diversifying selection refers to the situation in which natural selection favours different genotypes in different subenvironments.

The efficiency of diversifying natural selection is quite apparent in circumstances in which populations living a short distance apart have become genetically differentiated. In one example, populations of bent grass grow on heaps of mining refuse heavily contaminated with metals such as lead and copper. The soil has become so contaminated that it is toxic to most plants, but the dense bent grass stands growing over these refuse heaps have been shown to possess genes that make them resistant to high concentrations of lead and copper. Nonresistant bent grass plants grow a few metres from the contaminated soil. Bent grasses reproduce primarily by cross-pollination, so that the resistant grass receives wind-borne pollen from the neighbouring nonresistant plants. Yet they maintain their genetic differentiation because nonresistant seedlings are unable to grow in the contaminated soil. In nearby uncontaminated soil, the nonresistant seedlings outgrow the resistant ones. The evolution of these resistant strains has taken place in the fewer than 400 years since the mines were first opened.

Protective morphologies and coloration exist in many animals as a defense against predators or as a cover against prey. Sometimes an organism mimics the appearance of a different one for protection. Diversifying selection often occurs in association with mimicry. A species of swallowtail butterfly, *Papilio dardanus,* is endemic in tropical and southern Africa. The males have yellow and black wings, with the characteristic tails in the second pair of wings. But the females in many localities are conspicuously different from the males; their wings lack tails and have colour patterns that vary from place to place. The explanation stems from the fact that *P. dardanus* can be eaten safely by birds. Many other butterfly species are noxious to birds, and so they are carefully avoided as food. In localities where *P. dardanus* coexists with noxious butterfly species, the *P. dardanus* females have evolved an appearance that mimics the noxious species. The birds confuse the mimics with their models and do not prey upon them. In different localities the females mimic different species; there are some areas where two or even three different female forms exist, each mimicking different noxious species. Diversifying selection has resulted in different phenotypes of *P. dardanus* as a protection from bird predators.

Sexual selection. Mutual attraction between the sexes is an important factor in reproduction. The males and fe-

From the experiments of Dr. H.B.D. Kettlewell, University of Oxford; photographs by John S. Haywood

Figure 5: Industrial melanism in the peppered moth, *Biston betularia.* A typical light-gray moth and a darkly pigmented variant rest on two oak trees. On the lichen-covered trunk (left), the light-gray moth is inconspicuous; it is quite conspicuous on the soot-covered tree (right).

males of many animal species are similar in size and shape except for the sexual organs and secondary sexual characteristics such as the breasts of female mammals. There are, however, species in which the sexes exhibit striking dimorphism. Particularly in birds and mammals, the males are often larger and stronger, more brightly coloured, or endowed with conspicuous adornments. But bright colours make animals more visible to predators; the long plumage of male peacocks and birds of paradise and the enormous antlers of aged male deer are cumbersome loads in the best of cases. Darwin knew that natural selection could not be expected to favour the evolution of disadvantageous traits, and he was able to offer a solution to this problem. He proposed that such traits arise by "sexual selection," which "depends not on a struggle for existence in relation to other organic beings or to external conditions, but on a struggle between the individuals of one sex, generally the males, for the possession of the other sex."

The concept of sexual selection as a special form of natural selection is easily explained. Other things being equal, organisms more proficient in securing mates have higher fitness. There are two general circumstances leading to sexual selection. One is the preference of one sex (often the females) for individuals of the other sex that exhibit certain traits; the other is increased strength (usually among the males) that yields greater success in securing mates.

Selection by females among competing males

The presence of a particular trait among the members of one sex can make them somehow more attractive to the opposite sex. This type of "sex appeal" has been experimentally demonstrated in all sorts of animals, from vinegar flies to pigeons, mice, dogs, and rhesus monkeys. When, for example, *Drosophila* flies, some with yellow bodies and others with the normal yellowish-gray pigmentation, are placed together, normal males are preferred over yellow males by females with either body colour.

Sexual selection can also come about because a trait—the antlers of a stag, for example—increases prowess in competition with members of the same sex. Stags, rams, and bulls use antlers or horns in contests of strength; a winning male usually secures more female mates. Therefore, sexual selection may lead to increased size and aggressiveness in males. Male baboons are more than twice as large as the females, and the behaviour of the docile females contrasts with that of the aggressive males. A similar dimorphism occurs in the northern sea lion, *Eumetopias jubata,* where males weigh about 1,000 kilograms, about three times as much as females. The males fight fiercely in their competition for females; large, battle-scarred males occupy their own rocky islets, each holding a harem of up to 20 females. In many mammals that live in packs, troops, or herds—such as wolves, horses, and buffalos—there usually is a hierarchy of dominance based on age and strength, with males that rank high in the hierarchy doing most of the mating.

Altruism

Kin selection. The apparent altruistic behaviour of many animals is, like some manifestations of sexual selection, a trait that at first seems incompatible with the theory of natural selection. Altruism is a form of behaviour that benefits other individuals at the expense of the one that performs the action; the fitness of the altruist is diminished by its behaviour, whereas individuals that act selfishly benefit from it at no cost to themselves. Accordingly it might be expected that natural selection would foster the development of selfish behaviour and eliminate altruism. This conclusion is not so compelling when it is noticed that the beneficiaries of altruistic behaviour are usually relatives. They all carry the same genes, including the genes that promote altruistic behaviour. Altruism may evolve by kin selection, which is simply a type of natural selection in which relatives are taken into consideration when evaluating an individual's fitness.

Natural selection favours genes that increase the reproductive success of their carriers, but it is not necessary that all individuals with a given genotype have higher reproductive success. It suffices that carriers of the genotype reproduce more successfully on the average than those possessing alternative genotypes. A parent shares half of its genes with each progeny, so a gene that promotes parental altruism is favoured by selection if the behaviour's cost to the parent is less than half of its average benefits to the progeny. Such a gene will be more likely to increase in frequency through the generations than an alternative gene that does not promote altruistic behaviour. Parental care is, therefore, a form of altruism readily explained by kin selection. The parent spends some energy caring for the progeny because it increases the reproductive success of the parent's genes.

Kin selection extends beyond the relationship between parents and their offspring. It facilitates the development of altruistic behaviour when the energy invested, or the risk incurred, by an individual is compensated in excess by the benefits ensuing to relatives. The closer the relationship between the beneficiaries and the altruist, and the greater the number of beneficiaries, the higher the risks and efforts warranted in the altruist. Individuals that live together in a herd or troop usually are related and often behave toward each other in this way. Adult zebras, for instance, will turn toward an attacking predator to protect the young in the herd, rather than flee to protect themselves.

Altruism also occurs among unrelated individuals when the behaviour is reciprocal and the altruist's costs are smaller than the benefits to the recipient. This reciprocal altruism is found in the mutual grooming of chimpanzees and other primates as they clean each other of lice and other pests. Another example appears in flocks of birds that post sentinels to warn of danger. A crow sitting in a tree watching for predators, while the rest of the flock forages, incurs a small loss by not feeding; but this is well compensated by the protection it receives when it itself forages and others of the flock stand guard.

Social behaviour among insects

A particularly valuable contribution of the theory of kin selection is its explanation of the evolution of social behaviour among ants, bees, wasps, and other social insects. In the honeybee, for example, the female workers build the hive, care for the young, and gather food, but they are sterile; the queen bee alone produces progeny. It would seem that the workers' behaviour would in no way be promoted or maintained by natural selection. Any genes causing such behaviour would seem likely to be eliminated from the population, because individuals exhibiting the behaviour do not increase their own reproductive success, but that of the queen. The situation is, however, more complex.

Queen bees produce some eggs that remain unfertilized and develop into males, or drones, having a mother but no father. Their main role in the colony is to engage in the "nuptial flight," during which one of them fertilizes the queen. Other eggs laid by queen bees are fertilized and develop into females, the large majority of which are workers. A queen typically mates with a single male once during her lifetime; the male's sperm is stored in the queen's spermatheca, from which it is gradually released as she lays fertilized eggs. All the queen's female progeny therefore have the same father, so that workers are more closely related to one another and to any new sister queen than they are to the mother queen. The female workers receive one-half of their genes from the mother and one-half from the father, but they share among themselves three-quarters of their genes. One half of their genes come from the father; this half of the set is the same in every worker, because the father had only one set of genes rather than two to pass on (the male developed from an unfertilized egg, so all his sperm carry the same set of genes). The other half of the workers' genes come from the mother, and on the average half of them are identical in any two sisters. Consequently, with three-quarters of her genes present in her sisters but only half of her genes able to be passed on to a daughter, a worker's genes are transmitted one-and-a-half times more effectively when she raises a sister (whether another worker or a new queen) than if she produces a daughter of her own.

Species and speciation

THE CONCEPT OF SPECIES

Darwin sought to explain the splendid multiformity of the living world: thousands of organisms of the most diverse kinds, from lowly worms to spectacular birds of paradise,

from yeasts and molds to oaks and orchids. His *Origin of Species* is a sustained argument showing that the diversity of organisms and their characteristics can be explained as the result of natural processes.

Species come about as the result of gradual change prompted by natural selection. Environments are continuously changing in time, and they differ from place to place. Natural selection, therefore, favours different characteristics in different situations. The accumulation of differences eventually yields different species.

Everyday experience teaches that there are different kinds of organisms and how to identify them. Everyone knows that people belong to the human species and are different from cats and dogs, which in turn are different from each other. There are differences among people, as well as among cats and dogs; but individuals of the same species are considerably more similar among themselves than they are to individuals of other species.

External similarity is the common basis for identifying individuals as being members of the same species. But there is more to it than that; a bulldog, a terrier, and a golden retriever are very different in appearance, but they are all dogs because they can interbreed. People can also interbreed with one another, and so can cats, but people cannot interbreed with dogs or cats, nor these with each other. It is, then, clear that although species are usually identified by appearance, there is something basic, of great biological significance, behind similarity of appearance; individuals of a species are able to interbreed with one another but not with members of other species. This is expressed in the following definition: Species are groups of interbreeding natural populations that are reproductively isolated from other such groups.

Ability of species members to interbreed

The ability to interbreed is of great evolutionary importance, because it determines that species are independent evolutionary units. Genetic changes originate in single individuals; they can spread by natural selection to all members of the species but not to individuals of other species. Individuals of a species share a common gene pool that is not shared by individuals of other species. Different species have independently evolving gene pools because they are reproductively isolated.

Although the criterion for deciding whether individuals belong to the same species is clear, there may be ambiguity in practice for two reasons. One is lack of knowledge; it may not be known for certain whether individuals living in different sites belong to the same species, because it is not known whether they can naturally interbreed. The other reason for ambiguity is rooted in the nature of evolution as a gradual process. Two geographically separate populations that at one time were members of the same species later may have diverged into two different species. Since the process is gradual, there is not a particular point at which it is possible to say that the two populations have become two different species.

A related situation pertains to organisms living at different times. There is no way to test whether or not today's humans could interbreed with those who lived thousands of years ago. It seems reasonable that living people, or living cats, would be able to interbreed with people, or cats, exactly like those that lived a few generations earlier. But what about the ancestors removed by 1,000 or 1,000,000 generations? The ancestors of modern humans that lived 500,000 years ago (about 20,000 generations) are classified in the species *Homo erectus,* whereas present-day humans are classified in a different species, *Homo sapiens.*

Relation of *Homo sapiens* to remote ancestors

There is not an exact time at which *Homo erectus* became *Homo sapiens,* but it would not be appropriate to classify remote human ancestors and modern humans in the same species just because the changes from one generation to the next are small. It is useful to distinguish between the two groups by means of different species names, just as it is useful to give different names to childhood and adulthood, even though there is no one moment when one passes from one to the other. Biologists distinguish species in organisms that lived at different times by means of a commonsense morphological criterion. If two organisms differ from each other about as much as two living individuals belonging to two different species differ, they will be classified in separate species and given different names.

The definition of species given above applies only to organisms able to interbreed. Bacteria and blue-green algae do not reproduce sexually, but by fission. Organisms that lack sexual reproduction are classified into different species according to criteria such as external morphology, chemical and physiological properties, and genetic constitution.

THE ORIGIN OF SPECIES

Reproductive isolation. In sexual organisms individuals able to interbreed belong to the same species. The biological properties of organisms that prevent interbreeding are called reproductive isolating mechanisms (RIM's). Oaks on different islands, minnows in different rivers, or squirrels in different mountain ranges cannot interbreed because they are physically separated, but not necessarily because they are biologically incompatible. Geographic separation, therefore, is not an RIM, since it is not a biological property of organisms.

There are two general categories of reproductive isolating mechanisms: prezygotic (those that take effect before fertilization) and postzygotic (those that take effect after). Prezygotic RIM's prevent the formation of hybrids between members of different populations through ecological, temporal, ethological (or behavioral), mechanical, and gametic isolation. Postzygotic RIM's reduce the viability or fertility of hybrids or their progeny.

Ecological isolation. Populations may occupy the same territory but live in different habitats and so not meet. The *Anopheles maculipennis* group consists of six mosquito species, some of which are involved in the transmission of malaria. Although the species are virtually indistinguishable morphologically, they are isolated reproductively, in part because they breed in different habitats. Some breed in brackish water, others in running fresh water, and still other in stagnant fresh water.

Temporal isolation. Populations may mate or flower at different seasons or different times of day. Three tropical orchid species of the genus *Dendrobium* flower for a single day; the flowers open at dawn and wither by nightfall. Flowering occurs in response to certain meteorological stimuli, such as a sudden storm on a hot day. The same stimulus acts on all three species, but the lapse between the stimulus and flowering is eight days in one species, nine in another, and 10 or 11 in the third. Interspecific fertilization becomes impossible because at the time when the flowers of one species open, those of the other species have already withered or are not yet mature.

A peculiar form of temporal isolation exists between pairs of closely related species of cicadas, in which one species of each pair emerges every 13 years, the other every 17 years. The two species of a pair may be sympatric (live in the same territory), but they have an opportunity to form hybrids only once every 221 (13×17) years.

Ethological (behavioral) isolation. Sexual attraction between males and females may be weak or absent. In most animal species, members of the two sexes must first search for each other and come together. Complex courtship rituals then take place, with the male often taking the initiative and the female responding. This in turn generates additional actions by the male and responses by the female, and eventually there is copulation (or, in the case of some aquatic organisms, release of the sex cells for fertilization in the water). These elaborate rituals are specific to a species and play a significant part in species recognition. If the sequence of events in the search-courting-mating process is rendered disharmonious by either of the two sexes, then the entire process will be interrupted. Courtship and mating rituals have been extensively analyzed in some mammals, birds, and fishes, and in a number of insect species.

Species-specific courtship rituals

Ethological isolation is often the most potent RIM to keep animal species from interbreeding. It can be remarkably strong even among closely related species. The vinegar flies *Drosophila serrata, D. birchii,* and *D. dominicana* are three sibling species (that is, they are nearly indistinguishable morphologically) that are endemic in Australia and on the islands of New Guinea and New Britain. In many

areas these three species occupy the same territory, but no hybrids are known to occur in nature. The strength of their ethological isolation has been tested in the laboratory by placing groups of 10 females of one strain and 10 males of another together for several days. In groups in which the two strains were of the same species but from different geographic origins, a large majority of the females (usually 90 percent or more) were fertilized; but no inseminations or very few (less than 4 percent) took place when males and females were of different species, whether from the same or different geographic origins.

It should be added that the rare interspecific inseminations that did occur among the vinegar flies produced hybrid adult individuals in very few instances, and the hybrids were always sterile. This illustrates a common pattern: reproductive isolation between species is achieved by several RIM's in succession; if one breaks down, others are still present. In addition to ethological isolation, hybrid inviability and hybrid sterility prevent successful breeding between members of the three *Drosophila* species and many other animal species as well.

Species recognition during courtship involves stimuli that may be chemical (olfactory), visual, auditory, or tactile. Pheromones are specific substances that play a critical role in recognition between members of a species; they have been chemically identified in ants, moths, butterflies, and mammals. The "songs" of birds, frogs, and insects (which produce them by vibrating or rubbing their wings) are species recognition signals. Some form of physical contact or touching occurs in many mammals, but also in *Drosophila* flies and other insects.

Mechanical isolation. Copulation is often impossible between different animal species because of incompatible shape and size of the genitalia; in plants, variations in flower structure may impede pollination. In two species of sage from California, the two-lipped flowers of *Salvia mellifera* have stamens and style in the upper lip, whereas *S. apiana* has long stamens and style and a specialized floral configuration. *S. mellifera* is pollinated by small or medium-sized bees that carry pollen on their backs from flower to flower. *S. apiana,* however, is pollinated by large carpenter bees and bumblebees that carry the pollen on their wings and other body parts. Even if the pollinators of one species visit flowers of the other, pollination cannot occur because the pollen does not come into contact with the style of the alternative species.

Floral defenses against interspecies pollination

Gametic isolation. Marine animals often discharge their eggs and spermatozoa into the surrounding water, where fertilization takes place. Gametes of different species may fail to attract one another. For example, the sea urchins *Strongylocentrotus purpuratus* and *S. franciscanus* can be induced to release their eggs and sperms simultaneously, but most of the fertilizations that result are between eggs and sperms of the same species. In animals with internal fertilization, spermatozoa may be unable to function in the sexual ducts of females of different species. In plants, pollen grains of one species typically fail to germinate on the stigma of another species, so that the pollen never reaches the ovary and fertilization cannot occur.

Hybrid inviability. Occasionally, prezygotic mechanisms are absent or break down so that interspecific zygotes are formed. These zygotes, however, often fail to develop into mature individuals. The hybrid embryos of sheep and goats, for example, die in the early developmental stages before birth. Hybrid inviability is common in plants, whose hybrid seeds often fail to germinate or die shortly after germination.

Hybrid sterility. Hybrid zygotes sometimes develop into adults, such as mules (hybrids between horses and donkeys), but the adults fail to develop functional gametes and are sterile.

Hybrid breakdown. In plants more than in animals, hybrids between closely related species are sometimes partially fertile. Gene exchange may nevertheless be inhibited because the offspring are poorly viable or sterile. Hybrids between the cotton species *Gossypium barbadense, G. hirsutum,* and *G. tomentosum* appear vigorous and fertile, but their progenies die in seed or early in development, or they develop into sparse, weak plants.

A model of speciation. Since species are groups of populations reproductively isolated from one another, asking about the origin of species is equivalent to asking how reproductive isolation arises between populations. Two theories have been advanced to answer this question. One theory considers isolation as an accidental by-product of genetic divergence. Populations that become genetically less and less alike (as a consequence, for example, of adaptation to different environments) may eventually be unable to interbreed because their gene pools are disharmonious. The other theory regards isolation as a product of natural selection. Whenever hybrid individuals are less fit than nonhybrids, natural selection will directly promote the development of RIM's. This occurs because genetic variants interfering with hybridization have greater fitness than those favouring hybridization, given that the latter are often present in poorly fit hybrids.

Two means of speciation

These two theories of the origin of reproductive isolation are not mutually exclusive. Reproductive isolation may indeed come about incidentally to genetic divergence between separated populations. Consider, for example, the evolution of many endemic species of plants and animals in the Hawaiian archipelago. The ancestors of these species arrived in the Hawaiian Islands several million years ago. There they evolved as they became adapted to the environmental conditions and colonizing opportunities found on the islands. Reproductive isolation between the populations evolving in Hawaii and the continental populations was never directly promoted by natural selection because their geographic remoteness forestalled any opportunities for hybridizing. Nevertheless, reproductive isolation became complete in many cases as a result of gradual genetic divergence over thousands of generations.

Frequently, however, the course of speciation involves the processes postulated by both theories; reproductive isolation starts as a by-product of gradual evolutionary divergence but is completed by natural selection directly promoting the evolution of prezygotic RIM's.

The two sets of processes identified by the two speciation theories may be seen, therefore, as two different stages in the splitting of one evolutionary lineage into two species. The process can start only when gene flow is somehow interrupted between two populations. Interruption may be due to geographic separation, or it may be initiated by some genetic change that affects some but not other individuals living in the same territory. Absence of gene flow makes it possible for the two populations to become genetically differentiated as a consequence of adapting to diverse local conditions and of genetic drift. It is necessary that gene flow be interrupted, because otherwise the two groups of individuals would still share in a common gene pool and fail to become genetically different. The two genetically isolated groups are likely to become more and more different as time goes on. Eventually some incipient reproductive isolation may take effect because the two gene pools are no longer coadapted. Hybrid individuals will carry genes combined from two gene pools and will therefore have reduced viability or fertility.

Separation of gene pools

The circumstances just described may persist for so long that the populations become completely differentiated into separate species. It happens quite commonly, however, in both animals and plants, that opportunities for hybridization arise between two populations that are becoming genetically differentiated. Two outcomes are possible. One is that the hybrids manifest little or no reduction of fitness, so that gene exchange between the two populations proceeds freely, eventually leading to their integration into a single gene pool. The second possible outcome is that reduction of fitness in the hybrids is sufficiently large for natural selection to favour the emergence of prezygotic RIM's preventing the formation of hybrids altogether. This situation may be identified as the second stage in the speciation process.

How natural selection brings about the evolution of prezygotic RIM's can be understood in the following way. Assume that there are gene variants in one of two populations, P1, that increase the probability that P1 individuals will choose P1 rather than P2 mates. Such gene variants will increase in frequency in the P1 population, because

they are more often present in the progenies of P1 × P1 matings, which have normal fitness. The alternative genetic variants that do not favour P1 × P1 matings will be more often present in the progenies of P1 × P2 matings, which have lower fitness. The same process will enhance the frequency in the P2 population of genetic variants that lead P2 individuals to choose P2 rather than P1 mates. Prezygotic RIM's may therefore evolve in both populations and lead to their becoming two separate species.

Two-stage speciation process

The two stages of the process of speciation can be characterized, finally, by outlining their distinctions. The first stage primarily involves the appearance of postzygotic RIM's as accidental by-products of overall genetic differentiation rather than as express targets of natural selection; the second stage involves the evolution of prezygotic RIM's that are directly promoted by natural selection. The first stage may come about suddenly, in one or a few generations, rather than as a long, gradual process. The second stage succeeds the first in time but need not always be present.

Geographic speciation. One common mode of speciation is known as geographic, or allopatric (in separate territories), speciation. The general model of the speciation process advanced in the previous paragraphs applies well to geographic speciation. The first stage begins as a result of geographic separation between populations. This may occur when a few colonizers reach a geographically separate habitat, perhaps an island, lake, river, isolated valley, or mountain range. In another process, a population may be split into two geographically separate ones by topographic changes, such as a cessation of water flow between two lakes, or by an invasion of competitors, parasites, or predators into the intermediate zone. If these types of geographic separation continue for some time, postzygotic RIM's may appear as a result of gradual genetic divergence.

In the second stage, an opportunity for interbreeding may later be brought about by topographic changes establishing continuity between the previously isolated territories or by ecological changes making the intermediate territory habitable for the organisms. If the fitness of hybrids of the formerly separated populations is sufficiently reduced, natural selection will foster the development of prezygotic RIM's, and the two populations may then evolve into two species.

Investigation has been made of many populations that are in the first stage of geographic speciation. There are fewer well-documented instances of the second stage, presumably because this occurs fairly rapidly (in evolutionary time).

Both stages of speciation are present in a group of six closely related species of New World *Drosophila* that have been extensively studied by evolutionists for more than three decades. Two of these sibling species, *D. willistoni* and *D. equinoxialis,* consist of groups of populations in the first stage of speciation and identified as different subspecies. Two *D. willistoni* subspecies live in continental South America; *D. w. quechua* lives west of the Andes and *D. w. willistoni* east of the Andes. They are effectively separated by the Andes because the flies cannot live at high altitudes. It is not known whether their geographic separation is as old as the Andes, but it has existed long enough for postzygotic RIM's to evolve. When the two subspecies are crossed in the laboratory, the hybrid males are completely sterile if the mother came from the *quechua* subspecies; but in the reciprocal cross all hybrids are fertile. If hybridization should occur in nature, selection would favour the evolution of prezygotic RIM's because of the complete sterility of half of the hybrid males.

Drosophila equinoxialis equinoxialis and *D. e. caribbensis* are another pair of subspecies. *D. e. equinoxialis* inhabits continental South America, and *D. e. caribbensis* lives in Central America and the Caribbean. Crosses made in the laboratory between these two subspecies always produce sterile males, irrespective of the subspecies of the mother. Natural selection would, then, promote prezygotic RIM's between these two subspecies more strongly than between those of *D. willistoni.* But laboratory experiments show no evidence of ethological isolation or any other prezygotic RIM, presumably because the geographic isolation of the subspecies has forestalled hybridization between members.

One more sibling species of the group is *Drosophila paulistorum,* a species that includes groups of populations well into the second stage of geographic speciation. Six such groups have been identified as semispecies, or incipient species, two or three of which are sympatric (occupying the same territory) in many localities. Male hybrids between individuals of the different semispecies are sterile; laboratory crosses always yield fertile females but sterile males.

Ongoing speciation of *Drosophila paulistorum*

Whenever two or three incipient species of *D. paulistorum* have come into contact, the second stage of speciation has led to the development of ethological isolation, which ranges from incipient to virtually complete. Laboratory experiments show that when both incipient species are from the same locality, their ethological isolation is complete, so that only individuals of the same incipient species mate. When the individuals from different incipient species come from different localities, however, ethological isolation is usually present but far from complete. This is precisely as the speciation model predicts. Natural selection effectively promotes ethological isolation where the incipient species are sympatric; but the genes responsible for this isolation have not yet fully spread to populations in which another semispecies is not present.

The eventual outcome of the process of geographic speciation is complete reproductive isolation as can be observed among the species of the *Drosophila* group. *D. willistoni, D. equinoxialis, D. tropicalis,* and *D. paulistorum* coexist sympatrically over wide regions of Central and South America while preserving their separate gene pools. Hybrids are not known in nature and are virtually impossible to obtain in the laboratory; and all males at least are completely sterile. This total reproductive isolation has evolved, however, with very little morphological differentiation. Females from different sibling species cannot be distinguished by experts; the males can be identified only by small differences in the shape of their genitalia, unrecognizable except under a microscope.

Adaptive radiation. The geographic separation of populations derived from common ancestors may continue long enough so that the populations become completely differentiated species before ever regaining sympatry. As the allopatric populations continue evolving independently, RIM's develop and morphological differences may arise. The second stage of speciation—with natural selection directly stimulating the evolution of RIM's—never comes about in such situations, because reproductive isolation takes place simply as a consequence of the continued separate evolution of the populations.

This form of allopatric speciation is particularly apparent when colonizers reach geographically remote areas, such as islands, where they find few or no competitors and have an opportunity to diverge as they become adapted to the new environment. Sometimes a multiplicity of new environments becomes available to the colonizers, giving rise to several different lineages and species. This process of rapid divergence of multiple species from a single ancestral lineage is called adaptive radiation.

Galápagos finches

Many examples of speciation by adaptive radiation are found in archipelagos removed from the mainland. The Galápagos Islands are about 600 miles off the west coast of South America. When Darwin arrived there in 1835, he discovered many species not found anywhere else in the world—for example, 14 species of finch (known as Galápagos, or Darwin's, finches). These passerine birds have adapted to a diversity of habitats and diets, some feeding mostly on plants, others exclusively on insects. The various shapes of their bills are clearly adapted to probing, grasping, biting, or crushing—the diverse ways in which the different Galápagos species obtain their food (Figure 6). The explanation for such diversity is that the ancestor of Galápagos finches arrived in the islands before other kinds of birds and encountered an abundance of unoccupied ecological niches. The finches underwent adaptive radiation, evolving a variety of species with ways of life capable of exploiting opportunities that in continental faunas are exploited by other species.

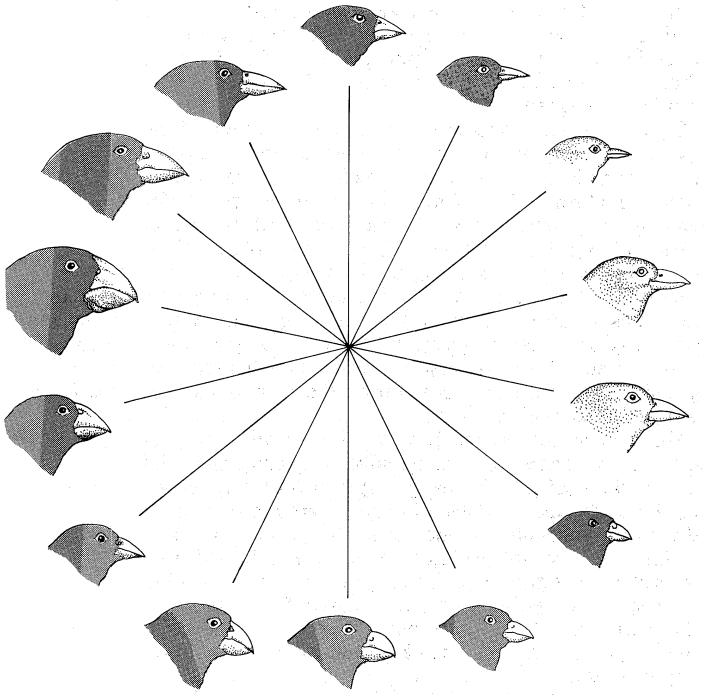


Figure 6: Fourteen species of Galápagos finches that evolved from a common ancestor. The different shapes of their bills, suited to different diets and habitats, show the process of adaptive radiation.
From P.R. Grant, *Ecology and Evolution of Darwin's Finches* (1986)

The Hawaiian Islands also provide striking examples of adaptive radiation. The archipelago consists of several volcanic islands, ranging from about 1,000,000 to more than 10,000,000 years in age, far away from any continent or even other large islands. In their relatively small total land area, an astounding number of plant and animal species exist. Most of the species have evolved in the islands. Among them about two dozen species (about one-third of them now extinct) of honeycreepers, birds of the family Drepanididae, all derived from a single immigrant form. In fact, all but one of Hawaii's 71 native bird species are endemic; that is, they have evolved there and are found nowhere else. More than 90 percent of the native species of flowering plants, land mollusks, and insects are also endemic, as are two-thirds of the 168 species of ferns.

There are more than 500 native Hawaiian species of *Drosophila* flies—about one-third of the world's total number of known species. Far greater morphological and ecological diversity exists among the species in Hawaii than anywhere else in the world. The species of *Drosophila* in Hawaii have diverged by adaptive radiation from one or a few colonizers, which encountered an assortment of ecological niches that in other lands were occupied by different groups of flies or insects but that were available for exploitation in these remote islands.

Quantum speciation. In some modes of speciation the first stage is achieved in a short period of time. These modes are known by a variety of names, such as "quantum," "rapid," and "saltational" speciation, all suggesting the shortening of time involved. They are also known as "sympatric" speciation, alluding to the fact that quantum speciation often leads to speciation between populations that exist in the same territory or habitat. An important form of quantum speciation, polyploidy, is discussed below.

Quantum speciation without polyploidy has been seen in the annual plant genus *Clarkia.* Two closely related species, *Clarkia biloba* and *C. lingulata,* are both native to California. *C. lingulata* is known only from two sites in the central Sierra Nevada at the southern periphery of the distribution of *C. biloba,* from which it evolved starting with translocations and other chromosomal mutations. Such chromosomal rearrangements arise suddenly but reduce the fertility of heterozygous individuals. *Clarkia* species are capable of self-fertilization, which facilitates the propagation of the chromosomal mutants in different sets of individuals even within a single locality. This makes hybridization possible with nonmutant individuals and allows the second stage of speciation to go ahead.

Chromosomal mutants

Chromosomal mutations are often the starting point of quantum speciation in animals, particularly in groups such as moles and other rodents that live underground or have little mobility. Mole rats of the group *Spalax ehrenbergi* in Israel and gophers of the group *Thomomys talpoides* in the northern Rocky Mountains are well-studied examples.

The speciation process may also be initiated by changes in just one or a few gene loci when these alterations result in a change of ecological niche or, in the case of parasites, a change of host. Many parasites use their host as a place for courtship and mating; so organisms with two different host preferences may become reproductively isolated. If the hybrids are poorly fit because they are not effective parasites in either of the two hosts, natural selection will favour the development of additional RIM's. This type of speciation seems to be common among parasitic insects, a large group comprising tens of thousands of species.

Polyploidy. The multiplication of entire sets of chromosomes is known as polyploidy. A diploid organism carries in the nucleus of each cell two sets of chromosomes, one inherited from each parent; a polyploid organism has three or more sets of chromosomes. Many cultivated plants are polyploid: bananas are triploid, potatoes are tetraploid, bread wheat is hexaploid, some strawberries are octaploid.

In animals, polyploidy is relatively rare because it disrupts the balance between the sex chromosome and the other chromosomes, a balance being required for the proper development of sex. Naturally polyploid species are found in hermaphroditic animals (individuals having both male and female organs), which include snails, earthworms, and planarians. They are also found in forms with parthenogenetic females (which produce viable progeny without fertilization), such as some beetles, sow bugs, goldfish, and salamanders.

All major groups of plants have polyploid species, but they are most common among flowering plants (angiosperms), of which about 47 percent are polyploids. Polyploidy is rare among gymnosperms, such as pines, firs, and cedars, although the redwood, *Sequoia sempervirens,* is a polyploid. Most polyploid plants are tetraploids. Polyploids with three, five, or some other odd-number multiple of the basic chromosome number are sterile, because the separation of homologous chromosomes cannot be achieved properly during formation of the sex cells. Some plants with an odd number of chromosome sets persist by means of asexual reproduction, particularly through human cultivation; the banana is one example.

Polyploidy is a mode of quantum speciation that yields the beginnings of a new species in just one or two generations. There are two kinds of polyploids: autopolyploids, which derive from a single species, and allopolyploids, which stem from a combination of chromosome sets from different species. Allopolyploid plant species are much more numerous than autopolyploids.

Origin of allopolyploids

An allopolyploid species can originate from two plant species that have the same diploid number of chromosomes. The chromosome complement of one species may be symbolized as *AA,* and the other *BB.* An interspecific hybrid, *AB,* will usually be sterile owing to abnormal chromosome pairing and segregation during formation of the gametes at meiosis. But chromosome doubling may occur as a consequence of abnormal mitosis, in which the chromosomes divide but the cell does not. In a hybrid, this results in a cell with four sets of chromosomes, *AABB.* Tetraploid plant cells may proliferate and produce branches and flowers. Because there are two chromosomes of each kind, functional diploid gametes with the constitution *AB* can be produced by the tetraploid flowers. The union of two such gametes at fertilization produces a tetraploid individual (*AABB*). In this way, self-fertilization in plants makes possible the formation of a tetraploid individual as the result of a single abnormal cell division.

Autopolyploids originate in a similar fashion, except that the individual in which the abnormal mitosis occurs is not

a hybrid. Self-fertilization thus enables a single individual to multiply and give rise to a population. This population is a new species, since polyploid individuals are reproductively isolated from their diploid ancestors. A cross between a tetraploid and a diploid yields triploid progeny, which are sterile.

GENETIC DIFFERENTIATION DURING SPECIATION

Genetic changes underlie all evolutionary processes. In order to understand speciation and its role in evolution, it is useful to know how much genetic change takes place during the course of species development. It is of considerable significance to ascertain whether new species arise by altering only a few genes, or whether the process requires drastic changes—a genetic "revolution," as postulated by some evolutionists in the past. The issue is best considered separately with respect to each of the two stages of speciation and to the various modes of speciation.

The question of how much genetic differentiation occurs during speciation has become answerable only in recent years, with the development of appropriate methods for comparing genes of different species. Genetic change is measured with two parameters: genetic identity (I), which estimates the proportion of genes that are identical in two populations; and genetic distance (D), which estimates the proportion of gene changes that have occurred in the separate evolution of two populations. The value of I may range between zero and one, which correspond to the extreme situations in which no or all genes are identical; D may range from zero to infinity. D can reach beyond unity because each gene may change more than once in one or both populations as evolution goes on for many generations.

Genetic identity and genetic distance

As a model of geographic speciation, the *Drosophila willistoni* group of flies offers the distinct advantage of exhibiting both stages of the speciation process. About 30 randomly selected genes have been studied in a large number of natural populations of these species. The results are summarized in Table 3. The most significant figures are those given in lines 2 and 3 of the Table, which represent the first and second stages, respectively, of the process of geographic speciation. $D = 0.230$ means that about 23 gene changes have occurred for every 100 gene loci in the separate evolution of two subspecies; that is, the sum of the changes that have occurred in the two separately evolving lineages is 23 percent of all the genes. These are populations well advanced in the first stage of speciation, as manifested by the sterility of the hybrid males.

The genetic distance between the incipient species is the same, within experimental error, as that between the subspecies, or 22.6 percent (line 3). This implies that the development of ethological isolation, as it is found in these populations, does not require many genetic changes beyond those that occurred during the first stage of speciation. Indeed, no additional gene changes were detected in these experiments. The absence of major genetic changes during the second stage of speciation can be understood by considering the role of natural selection, which directly promotes the evolution of prezygotic RIM's during the second stage, so that only genes modifying mate choice need to change. In contrast, the development of postzygotic RIM's during the first stage occurs only after there is substantial genetic differentiation between populations, because it comes about only as an incidental outcome of overall genetic divergence.

Sibling species, such as *D. willistoni* and *D. equinoxialis*, exhibit 58 gene changes for every 100 gene loci after their divergence from a common ancestor (line 4). It is noteworthy that this much genetic evolution has occurred without altering the external morphology of these organisms. In the evolution of morphologically different species (line 5), the number of gene changes is greater yet, as would be expected.

Table 3: Genetic Differentiation Between Populations of Drosophila Willistoni

level of comparison	I (genetic identity)	D (genetic distance)
1. Local populations	0.970 ± 0.006	0.031 ± 0.007
2. Subspecies	0.795 ± 0.013	0.230 ± 0.016
3. Incipient species	0.798 ± 0.026	0.226 ± 0.033
4. Sibling species	0.563 ± 0.023	0.581 ± 0.039
5. Morphologically different species	0.352 ± 0.023	1.056 ± 0.068

Genetic changes concomitant with one or other of the two stages in the speciation process have been studied in a number of organisms, from insects and other invertebrates to all sorts of vertebrates, including mammals. The amount of genetic change during geographic speciation varies among organisms, but the two main observations made in the *D. willistoni* group seem to apply quite generally. These are that the evolution of postzygotic mechanisms during the first stage is accompanied by substantial genetic change (a majority of values range between $D = 0.15$ and $D = 0.30$) and that relatively few additional genetic changes are required during the second stage of geographic speciation.

Genetic change in the first stage of speciation

The conclusions drawn from the investigation of geographic speciation make it possible to predict the relative amounts of genetic change expected in the quantum modes of speciation. Polyploid species are a special case; they arise suddenly in one or a few generations, and at first they are not expected to be genetically different from their ancestors. More generally, quantum speciation involves a shortening of the first stage of speciation, so that postzygotic RIM's arise directly as a consequence of specific genetic changes (such as chromosome mutations). Populations in the first stage of quantum speciation, therefore, need not be substantially different in individual gene loci. This has been confirmed by genetic investigations of species recently arisen by quantum speciation. For example, the average genetic distance between four incipient species of the mole rat *S. ehrenbergi* is $D = 0.022$, and between those of the gopher *T. talpoides* it is $D = 0.078$. The second stage of speciation is modulated in essentially the same way as in the geographic mode. Not many gene changes are needed in either case to complete speciation.

Patterns and rates of species evolution

RECONSTRUCTION OF EVOLUTIONARY HISTORY

Evolution within a lineage and by lineage splitting. Evolution can take place by anagenesis, in which changes occur within a lineage; or by cladogenesis, in which a lineage splits into two or more separate lines. Anagenetic evolution has, over the course of 2,000,000 years, doubled the size of the human cranium; in the lineage of the horse, it has reduced the number of toes from four to one. Cladogenetic evolution has produced the extraordinary diversity of the living world, with its more than 2,000,000 species of animals, plants, fungi, and microorganisms.

The most essential cladogenetic function is speciation, the process by which one species splits into two or more species. Because species are reproductively isolated from one another, they are independent evolutionary units; that is, evolutionary changes occurring in one species are not shared with other species. Over time, species become more and more divergent from one another as a consequence of anagenetic evolution. Descendant lineages of two related species that existed millions of years ago may now be classified into quite different taxonomic categories, such as different genera or even different families.

The evolution of all living organisms, or of a subset of them, can be seen as a tree, with branches that divide into two or more as time progresses. Such "trees" are called phylogenies. Their branches represent evolving lineages, some of which eventually die out, while others persist in themselves or in their derived lineages down to the present time. Evolutionists are interested in the history of life and hence in the topology, or configuration, of phylogenies. They are concerned as well with the nature of the anagenetic changes along lineages and with the timing of the events.

Phylogenies

Phylogenetic relationships are ascertained by means of several complementary sources of evidence. First, there are the discovered remnants of organisms that lived in the past, the fossil record, which provides definitive evidence of relationships among some groups of organisms. The

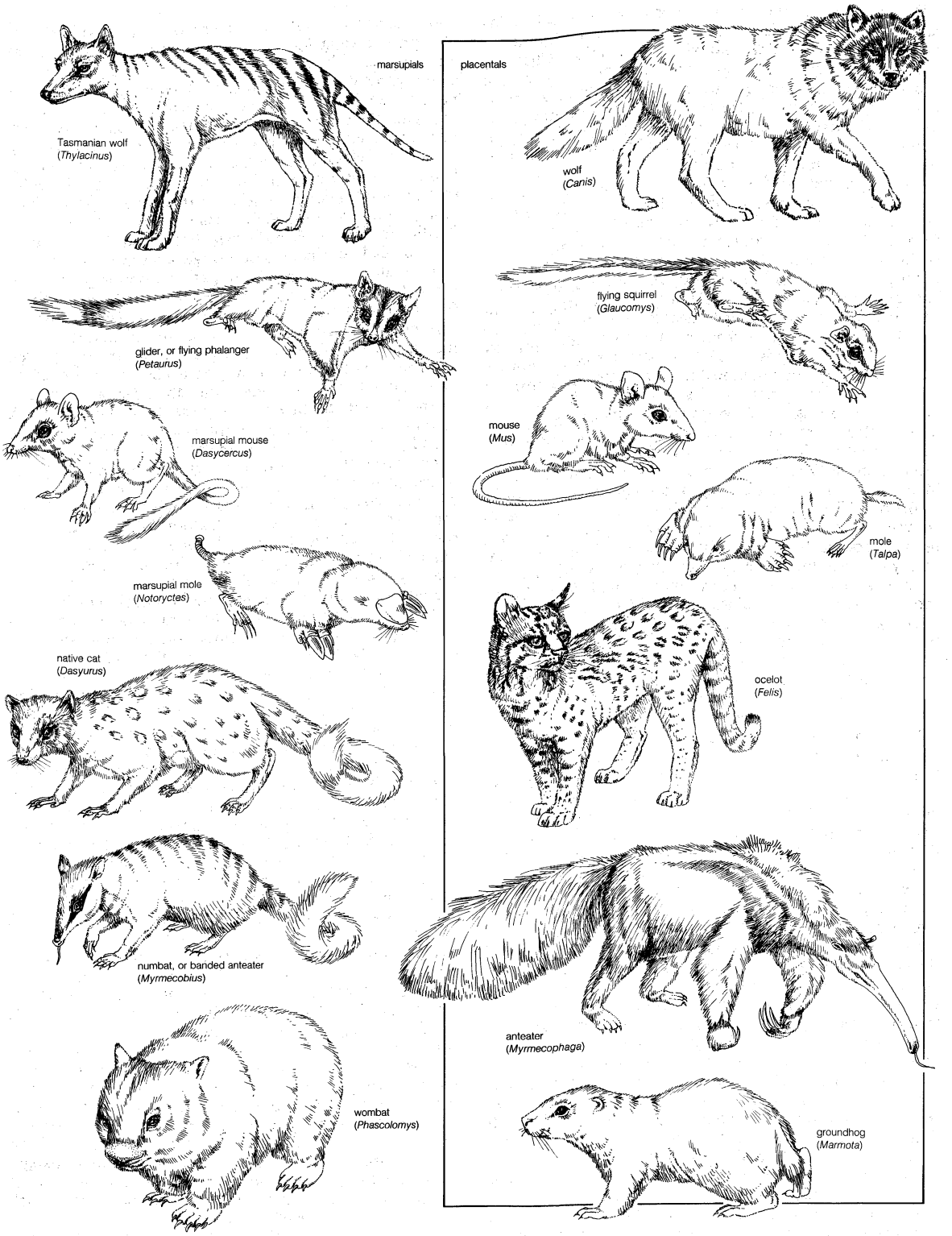


Figure 7: Parallel evolution of marsupial mammals in Australia and placental mammals on other continents.

fossil record, however, is far from complete and is often seriously deficient. Second, information about phylogeny comes from comparative studies of living forms. Comparative anatomy contributed the most information in the past, although additional knowledge came from comparative embryology, cytology, ethology, biogeography, and other biological disciplines. In recent years the comparative study of informational macromolecules—proteins and nucleic acids—has become a powerful tool for the study of phylogeny.

Morphological similarities among organisms have probably always been recognized. In ancient times, Aristotle and later his followers and those of Plato, particularly Porphyry, classified organisms (as well as inanimate objects) on the basis of similarities. The Aristotelian system of classification was further developed by some medieval Scholastics, notably Albertus Magnus and Thomas Aquinas. The modern foundations of taxonomy, the science of classification, were laid in the 18th century by Linnaeus and by the French botanist Michel Adanson. Lamarck dedicated much of his work to the systematic classification of organisms. He proposed that their similarities were due to ancestral relationships—in other words, to the degree of evolutionary proximity.

The modern theory of evolution provides a causal explanation of the similarities among living things. Organisms evolve by a process of descent with modification. Changes, and therefore differences, gradually accumulate over the generations. The more recent the last common ancestor of a group of organisms, the less their differentiation; similarities of form and function reflect phylogenetic propinquity. Accordingly, phylogenetic affinities can be inferred on the basis of relative similarity.

Convergent and parallel evolution. A distinction has to be made between resemblances due to propinquity of descent and those due only to similarity of function.

Homology and analogy

Correspondence of features in different organisms that is due to inheritance from a common ancestor is called homology. The forelimbs of humans, whales, dogs, and bats are homologous. The skeletons of these limbs are all constructed of bones arranged according to the same pattern because they derive from an ancestor with similarly arranged forelimbs (Figure 3). Correspondence of features due to similarity of function but not related to common descent is termed analogy. The wings of birds and of flies are analogous. Their wings are not modified versions of a structure present in a common ancestor but rather have developed independently as adaptations to a common function, flying. The similarities between the wings of bats and birds are partially homologous and partially analogous. The skeletal structure is homologous, owing to common descent from the forelimb of a reptilian ancestor; but the modifications for flying are different and independently evolved, and in this respect they are analogous.

Features that become more rather than less similar through independent evolution are said to be convergent. Convergence is often associated with similarity of function, as in the evolution of wings in birds, bats, and flies. The shark (a fish) and the dolphin (a mammal) are much alike in external morphology; their similarities are due to convergence, since they have evolved independently as adaptations to aquatic life.

Taxonomists also speak of parallel evolution. Parallelism and convergence are not always clearly distinguishable. Strictly speaking, convergent evolution occurs when descendants resemble each other more than their ancestors did with respect to some feature. Parallel evolution implies that two or more lineages have changed in similar ways, so that the evolved descendants are as similar to each other as their ancestors were. The evolution of marsupials in Australia paralleled the evolution of placental mammals in other parts of the world. There are Australian marsupials resembling true wolves, cats, mice, squirrels, moles, groundhogs, and anteaters. These placental mammals and the corresponding Australian marsupials evolved independently but in parallel lines by reason of their adaptation to similar ways of life. Some resemblances between a true anteater (*Myrmecophaga*) and a marsupial anteater (*Myrmecobius*) are due to homology—both are mammals. Others are due to analogy—both feed on ants.

Parallel and convergent evolution are also common in plants. New World cacti and African euphorbias are alike in overall appearance although they belong to separate families. Both are succulent, spiny, water-storing plants adapted to the arid conditions of the desert. Their corresponding morphologies have evolved independently in response to similar environmental challenges.

Serial homology

Homology can be recognized not only between different organisms but also between repetitive structures of the same organism. This has been called serial homology. There is serial homology, for example, between the arms and legs of humans, among the seven cervical vertebrae of mammals, and among the branches or leaves of a tree. The jointed appendages of arthropods are elaborate examples of serial homology. Crayfish have 19 pairs of appendages, all built according to the same basic pattern but serving diverse functions—sensing, chewing, food handling, walking, mating, egg carrying, and swimming. Serial homologies are not useful in reconstructing the phylogenetic relationships of organisms, but they are an important dimension of the evolutionary process.

Relationships in some sense akin to those between serial homologs exist at the molecular level between genes and proteins derived from ancestral gene duplications. The genes coding for the various hemoglobin chains are an example. About 500,000,000 years ago a chromosome segment carrying the gene coding for hemoglobin became duplicated, so that the genes in the different segments thereafter evolved in somewhat different ways, one eventually giving rise to the modern gene coding for α hemoglobin, the other for β hemoglobin. The β hemoglobin gene became duplicated again about 200,000,000 years ago, giving rise to the γ (fetal) hemoglobin. The α, β, γ, and other hemoglobin genes are homologous; similarities in their nucleotide sequences occur because they are modified descendants of a single ancestral sequence.

There are two ways of comparing homology between hemoglobins. One is to compare the same hemoglobin—for instance, the α chain—in different species of animals. The degree of divergence between the α chains reflects the degree of the evolutionary relationship among the organisms, because the hemoglobin chains have evolved independently of one another since the time of divergence of the lineages leading to the present-day organisms. A second way is to make comparisons between, say, the α and β hemoglobins of a single species. The degree of divergence between the different globin chains reflects the degree of relationship among the genes coding for them. The different globins have evolved independently of each other since the time of duplication of their ancestral genes. Comparisons between homologous genes or proteins within a given organism provide information about the phylogenetic history of the genes and, hence, about the historical sequence of the gene duplication events.

Whether similar features in different organisms are homologous or analogous—or simply accidental—cannot always be decided unambiguously, but the distinction must be made in order to determine phylogenetic relationships. Moreover, the degrees of homology must be quantified in some way so as to determine the propinquity of common descent among species. Difficulties arise here as well. Consider the forelimbs shown in Figure 3. It is not clear whether the homologies are greater between man and bird than between man and reptile, or between man and reptile than between man and bat. The fossil record sometimes provides the appropriate information, even though the record is deficient. Fossil evidence must be examined together with the evidence from comparative studies of living forms and with the quantitative estimates provided by comparative studies of proteins and nucleic acids.

Gradual and punctuational evolution. The fossil record indicates that morphological evolution is by and large a gradual process. Major evolutionary changes are usually due to a building up over the ages of relatively small changes. But the fossil record is discontinuous. Fossil strata are separated by sharp boundaries; accumulation of fossils within a geologic deposit (stratum) is fairly constant over time, but the transition from one stratum to another may involve gaps of tens of thousands of years. New species, characterized by small but discontinuous morphological changes, typically appear at the boundaries between strata, whereas the fossils within a stratum exhibit little morphological variation. That is not to say that the transition from one stratum to another always involves sudden changes in morphology; on the contrary, fossil forms often persist virtually unchanged through several geologic strata, each representing millions of years.

Gaps in the fossil record

The apparent morphological discontinuities of the fossil record are often attributed by paleontologists to the discontinuity of the sediments; that is, to the substantial time gaps encompassed in the boundaries between strata. The assumption is that, if the fossil deposits were more continuous, they would show a more gradual transition of form. Even so, morphological evolution would not always keep progressing gradually, because some forms, at least, remain unchanged for extremely long times. Examples are the lineages known as "living fossils": the lamp shell *Lingula*, a genus of brachiopod that appears to have remained essentially unchanged since the Ordovician Period, some 450,000,000 years ago; or the tuatara (*Sphenodon punctatus*), a reptile that has shown little morphological evolution for nearly 200,000,000 years since the early Mesozoic.

Some paleontologists have proposed that the discontinuities of the fossil record are not artifacts created by gaps in the record, but rather reflect the true nature of morphological evolution, which happens in sudden bursts associated with the formation of new species. The lack of morphological evolution, or stasis, of lineages such as *Lingula* and *Sphenodon* is in turn due to lack of speciation within those lineages. The proposition that morphological evolution is jerky, with most morphological change occurring during the brief speciation events and virtually no change during the subsequent existence of the species, is known as the punctuated equilibrium model of morphological evolution.

Whether morphological evolution in the fossil record is predominantly punctuational or gradual is a much debated question. The imperfection of the record makes it unlikely that the issue will be settled in the foreseeable future. Intensive study of a favourable and abundant set of fossils may be expected to substantiate punctuated or gradual evolution in particular cases. But the argument is not about whether only one or the other pattern ever occurs; it is about their relative frequency. Some paleontologists argue that morphological evolution is in most cases gradual and only rarely jerky, whereas others think the opposite is true.

Much of the problem is that gradualness or jerkiness is in the eye of the beholder. Consider the evolution of rib strength (the ratio of rib height to rib width) within a lineage of fossil brachiopods of the genus *Eocelia*. An abundant sample of fossils from the Silurian Period in Wales has been analyzed, with the results shown in Figure 8.

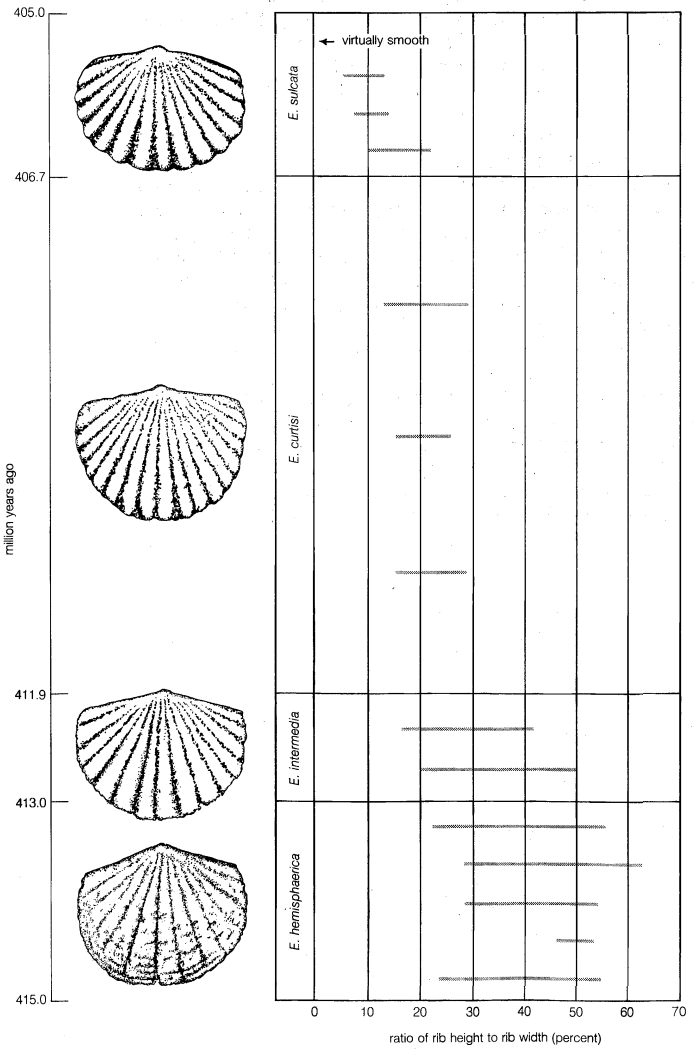


Figure 8: Rib strength in the evolution of the brachiopod *Eocelia*. The horizontal bars indicate the observed range of rib strength among fossilized finds.

One possible interpretation of the data is that rib strength changed little or not at all from 415,000,000 to 413,000,000 years ago; rapid change ensued for the next 1,000,000 years, with virtual stasis from 412,000,000 to 407,000,000 years ago; another short burst of change occurred around 406,000,000 years ago, followed by a final period of stasis. On the other hand, the record shown in the Figure may be interpreted as not particularly punctuated but rather as a gradual process, with the rate of change somewhat greater at particular times.

Problems of interpreting fossil data

The proponents of the punctuated equilibrium model propose not only that morphological evolution is jerky but also that it is associated with speciation events. They argue that phyletic evolution—that is, evolution along lineages of descent—proceeds at two levels. First, there is continuous change through time within a population. This consists largely of gene substitutions prompted by natural selection, mutation, genetic drift, and other genetic processes that operate at the level of the individual organism. The punctualists maintain that this continuous evolution within established lineages rarely, if ever, yields substantial morphological changes in species. Second, they say, there is the process of origination and extinction of species, in which most morphological change occurs. According to the punctualist model, evolutionary trends result from the patterns of origination and extinction of species rather than from evolution within established lineages.

Species are groups of interbreeding natural populations that are reproductively isolated from any other such groups. Speciation involves, therefore, the development of reproductive isolation between populations previously able to interbreed. Paleontologists recognize species by their different morphologies as preserved in the fossil record, but fossils cannot provide evidence of the development of reproductive isolation because new species that are reproductively isolated from their ancestors are often morphologically indistinguishable from them. Speciation as seen by paleontologists always involves substantial morphological change because paleontologists identify new species by morphological differences. This situation creates an insuperable difficulty for resolving the question whether morphological evolution is always associated with speciation events. If speciation is defined as the evolution of reproductive isolation, the fossil record provides no evidence of a necessary association between speciation and morphological change. But if new species are identified in the fossil record by morphological changes, then all such changes will occur concomitantly with the origination of new species.

Speciation and morphological change

MOLECULAR EVOLUTION

DNA and protein as informational macromolecules. The advances of molecular biology have made possible the comparative study of proteins and the nucleic acids, DNA and RNA. The DNA is the repository of hereditary (evolutionary and developmental) information. The relationship of proteins to the DNA is so immediate that they closely reflect the hereditary information. This reflection is not perfect, because the genetic code is redundant and, consequently, some differences in the DNA do not yield differences in the proteins. Moreover, it is not complete, because a large fraction of the DNA (about 90 percent in many organisms) does not code for proteins. Nevertheless, proteins are so closely related to the information contained in the DNA that they, as well as the nucleic acids, are called informational macromolecules.

Nucleic acids and proteins are linear molecules made up of sequences of units—nucleotides in the case of nucleic acids, amino acids in the case of proteins—which retain considerable amounts of evolutionary information. Comparing two macromolecules establishes the number of their units that are different. Because evolution usually occurs by changing one unit at a time, the number of differences is an indication of the recency of common ancestry. Changes in evolutionary rates may create difficulties, but macromolecular studies have two notable advantages over comparative anatomy and the other classical disciplines. One is that the information is more readily quantifiable. The number of units that are different is readily estab-

lished when the sequence of units is known for a given macromolecule in different organisms. The other advantage is that comparisons can be made even between very different sorts of organisms. There is very little that comparative anatomy can say when organisms as diverse as yeasts, pine trees, and human beings are compared; but there are homologous macromolecules that can be compared in all three.

Molecular information about lineage history

Informational macromolecules provide information not only about the topology of evolutionary history (cladogenesis) but also about the amount of genetic change that has occurred in any given lineage (anagenesis). It might seem at first that quantifying anagenesis for proteins and nucleic acids would be impossible, because it would require comparison of molecules from organisms that lived in the past with those from living organisms. Organisms of the past are sometimes preserved as fossils, but their DNA and proteins have largely disintegrated. Nevertheless, comparisons between living species provide information about anagenesis.

The following is an example of such comparison: Two living species, C and D, have a common ancestor, the extinct species B (Figure 9). If C and D were found to differ

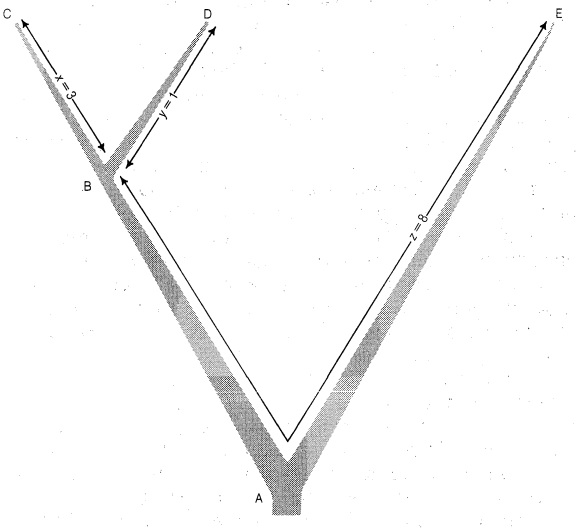


Figure 9: Amount of change in the evolutionary history of three living species (C, D, and E), inferred by comparing amino-acid differences in their myoglobin molecules.

by four amino-acid substitutions in a single protein, then it could safely be assumed that two substitutions (four total changes divided by two species) had taken place in the evolutionary lineage of each species. This assumption, however, could be invalidated by the discovery of a third living species, E, that is related to C, D, and their ancestor, B, through an earlier ancestor, A. The number of amino-acid differences between the protein molecules of the three living species may be as follows:

C and D = 4
C and E = 11
D and E = 9

Figure 9 proposes a phylogeny of the three living species, making it possible to estimate the number of amino-acid substitutions that have occurred in each lineage. Let x denote the number of differences between B and C, y denote the differences between B and D, and z denote the differences between A and B as well as A and E. The following three equations can be produced:

$$x + y = 4$$
$$x + z = 11$$
$$y + z = 9$$

Solving the equations yields $x = 3$, $y = 1$, and $z = 8$.

As a concrete example, consider cytochrome c, a protein involved in cell respiration. The sequence of amino acids in this protein is known for many organisms, from bacteria and yeast to insects and humans; in animals, cytochrome c consists of 104 amino acids. When the amino-acid sequences of humans and rhesus monkeys are compared, they are found to be different at position 66 (isoleucine in humans, threonine in rhesus monkeys), but identical at the other 103 positions. When humans are compared with horses, 12 amino-acid differences are found; but when horses are compared with rhesus monkeys there are only 11 amino-acid differences. Even without knowing anything else about the evolutionary history of mammals, one would conclude that the lineages of humans and rhesus monkeys diverged from each other much more recently than they diverged from the horse lineage. Moreover, it can be concluded that the amino-acid difference between humans and rhesus monkeys must have occurred in the human lineage after its separation from the rhesus monkey lineage (see Figure 10).

Molecular phylogenies of species. Protein sequencing is one of several molecular methods developed for estimating genetic change during evolution. The effectiveness of this method can be illustrated by again using as an example the protein cytochrome c, whose amino-acid sequences are well known. Phylogenies can be constructed based on the number of amino-acid differences between species. But the amino-acid sequence of a protein contains more information than is reflected in the number of amino-acid differences. This is because the replacement of one amino acid by another in some cases requires no more than one nucleotide substitution in the DNA that codes for the protein but, in other cases, requires at least two nucleotide changes. Table 4 shows the minimum number of nucleotide differences in the genes of 20 separate species that are necessary to account for the amino-acid differences in their cytochrome c. Figure 11 proposes a phylogenetic tree based on the data in Table 4, showing the minimum numbers of nucleotide changes in each branch.

The relationships between species as shown in Figure 11 correspond fairly well with the relationships determined from other sources, such as the fossil record. According to Table 4, chickens are less closely related to ducks and pigeons than to penguins, and humans and monkeys diverged from the other mammals before the marsupial kangaroo separated from the nonprimate placentals. These are known to be erroneous relationships; but the power of the method is apparent in that a single protein yields a fairly accurate reconstruction of the evolutionary history of 20 organisms that started to diverge more than 1,000,000,000 years ago.

Cytochrome c is a slowly evolving protein. Widely different species have in common a large proportion of the amino acids in their cytochrome c, making possible the study of genetic differences among organisms only remotely related. For the same reason, however, comparing cytochrome c cannot determine evolutionary change in closely related species. For example, the amino-acid sequence of cytochrome c in humans and chimpanzees is

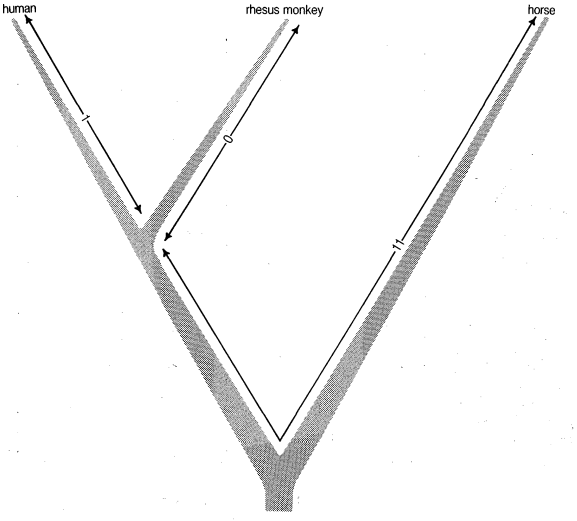


Figure 10: Phylogeny of the human, rhesus monkey, and horse, based on amino-acid substitutions in the evolution of cytochrome c in the lineages of the three species.

Table 4: Minimum Number of Nucleotide Differences in Genes Coding for Cytochrome c in 20 Different Organisms

organism	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Human	—	1	13	17	16	13	12	12	17	16	18	18	19	20	31	33	36	63	56	66
2. Monkey			12	16	15	12	11	13	16	15	17	17	18	21	32	32	35	62	57	65
3. Dog				10	8	4	6	7	12	12	14	14	13	30	29	24	28	64	61	66
4. Horse					1	5	11	11	16	16	16	17	16	32	27	24	33	64	60	68
5. Donkey						4	10	12	15	15	15	16	15	31	26	25	32	64	59	67
6. Pig							6	7	13	13	13	14	13	30	25	26	31	64	59	67
7. Rabbit								7	10	8	11	11	11	25	26	23	29	62	59	67
8. Kangaroo									14	14	15	13	14	30	27	26	31	66	58	68
9. Duck										3	3	3	7	24	26	25	29	61	62	66
10. Pigeon											4	4	8	24	27	26	30	59	62	66
11. Chicken												2	8	28	26	26	31	61	62	66
12. Penguin													8	28	27	28	30	62	61	65
13. Turtle														30	27	30	33	65	64	67
14. Rattlesnake															38	40	41	61	61	69
15. Tuna																34	41	72	66	69
16. Screwworm																	16	58	63	65
17. Moth																		59	60	61
18. *Neurospora* (mold)																			57	61
19. *Saccharomyces* (yeast)																				41
20. *Candida* (yeast)																				—

Source: Walter M. Fitch, *Science*, vol. 155, Jan. 20, 1967, p. 281, copyright 1967 by the AAAS.

identical, although they diverged about 10,000,000 years ago; between humans and rhesus monkeys, who diverged from their common ancestor 50,000,000 to 40,000,000 years ago, it differs by only one amino-acid replacement.

Other proteins that evolve more rapidly can be studied in order to establish phylogenetic relationships among closely related species. Genetic changes in the evolution of such species can also be studied by DNA sequencing, DNA hybridization, immunology, and gel electrophoresis.

Molecular phylogenies of genes. It is now possible to obtain the nucleotide sequence of the DNA. Although the number of genes that have been sequenced is relatively small, new sequences are being worked out at a fast rate. The use of DNA sequences in evolutionary research has so far been rewarding, particularly in the study of gene duplications. The genes that code for the hemoglobins in humans and a few other mammals provide the best example.

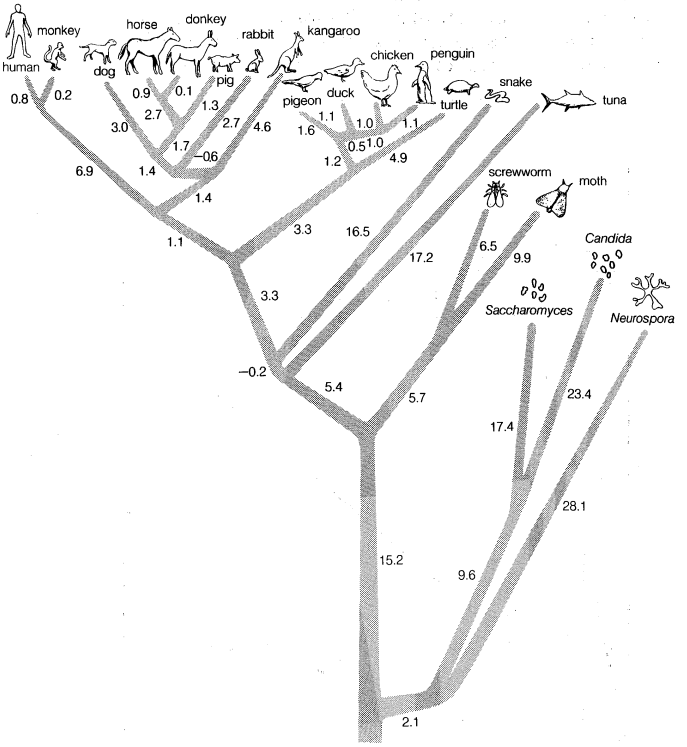


Figure 11: Phylogeny based on differences in the protein sequence of cytochrome c in organisms ranging from *Neurospora* mold to humans. The numbers estimate the nucleotide substitutions that have occurred along the lineages in the gene coding for this protein.

Hemoglobin studies

The amino-acid sequences of the hemoglobin chains and of myoglobin, a closely related protein, are known for humans as well as for other organisms. These sequences have made it possible to reconstruct the evolutionary history of the duplications that gave rise to the corresponding genes. But direct examination of the nucleotide sequences of the coding for these proteins has shown that the situation is more complex, and also more interesting, than it appears from the protein sequences.

The DNA sequence studies on human hemoglobin genes have shown that their number is greater than previously thought. The hemoglobins are tetramers, consisting of two polypeptides of one kind and two of another kind. One of the two kinds of polypeptide is ε in embryonic hemoglobin, γ in fetal hemoglobin, β in adult hemoglobin A, and δ in adult hemoglobin A_2. (Hemoglobin A makes up about 98 percent, and hemoglobin A_2 about 2 percent, of human adult hemoglobin). The other kind of polypeptide is ζ in embryonic hemoglobin and α in fetal and adult hemoglobin. The genes coding for one kind of polypeptide (ε, γ, β, and δ) are located on chromosome 11; the genes coding for the second kind of polypeptide (ζ and α) are located on chromosome 16.

But there are additional complexities. Two γ genes exist, known as G_γ and A_γ, as do two α genes (α_1 and α_2). Furthermore, there are two β pseudogenes ($\psi\beta_1$ and $\psi\beta_2$) and two α pseudogenes ($\psi\alpha_1$ and $\psi\alpha_2$), as well as a pseudo-ζ. These pseudogenes are very similar in nucleotide sequence to the corresponding functional genes, but they include terminating codons and other mutations that make it impossible for them to yield functional hemoglobins.

The similarity in the nucleotide sequence of the globin genes, and pseudogenes, of both the α and β gene families indicates that they are all homologous; that is, that they have arisen through various duplications and subsequent evolution from a gene ancestral to all. Moreover, homology also exists between the nucleotide sequences that separate one gene from another. The evolutionary history of the globin genes is summarized in Figure 12.

The molecular clock of evolution. One conspicuous attribute of molecular evolution is that differences between homologous molecules can readily be quantified and expressed, as, for example, proportions of nucleotides or amino acids that have changed. Rates of evolutionary change can, therefore, be more precisely established with respect to DNA or proteins than with respect to phenotypic traits of form and function. Studies of molecular evolution rates have led to the proposition that macromolecules may serve as evolutionary clocks.

Change in amino-acid sequence over time

It was first observed in the 1960s that the numbers of amino-acid differences between homologous proteins of any two given species seemed to be nearly proportional to the time of their divergence from a common ancestor. If the rate of evolution of a protein or gene were approximately the same in the evolutionary lineages leading

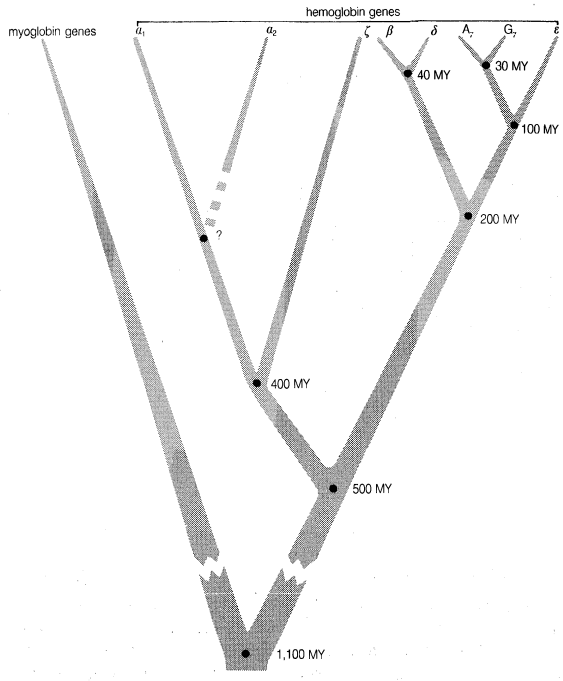


Figure 12: Evolutionary history of the globin genes. The dots indicate points at which ancestral genes duplicated, giving rise to new gene lineages. The approximate times when these duplications occurred are indicated in millions of years (MY) ago. The time when the duplication of α_1 and α_2 occurred is uncertain.

to different species, proteins and DNA sequences would provide a molecular clock of evolution. The sequences could then be used not only to reconstruct the topology of a phylogeny—that is, the sequence of branching events—but also the time when the various events occurred.

Consider, for example, the phylogeny shown in Figure 11. If the substitution of nucleotides in the gene coding for cytochrome c occurred at a constant rate through time, one could determine the time elapsed along any branch of the phylogeny simply by examining the number of nucleotide substitutions along that branch. One would need only to calibrate the clock by reference to an outside source, such as the fossil record, that would give us the actual geologic time elapsed in at least one specific lineage.

The molecular evolutionary clock, of course, is not expected to be a metronomic clock, like a watch or other timepiece that measures time exactly, but a stochastic clock like radioactive decay. In a stochastic clock, the probability of a certain amount of change is constant, although some variation occurs in the actual amount of change. Over fairly long periods of time a stochastic clock is quite accurate. The enormous potential of the molecular evolutionary clock lies in the fact that each gene or protein is a separate clock. Each clock "ticks" at a different rate—the rate of evolution characteristic of a particular gene or protein (see Figure 13)—but each of the thousands and thousands of genes or proteins provides an independent measure of the same evolutionary events.

Evolutionists have found that the amount of variation observed in the evolution of DNA and proteins is greater than is expected from a stochastic clock; in other words, the clock is erratic. The discrepancies in evolutionary rates along different lineages are not excessively large, however. So it is possible, in principle, to time phylogenetic events with as much accuracy as may be desired; but more genes or proteins (about two to four times as many) must be examined than would be required if the clock were stochastically constant. The average rates obtained for several proteins taken together become a fairly precise clock, particularly when many species are studied and the evolutionary events involve long time periods (on the order of 50,000,000 years or more).

This conclusion is illustrated in Figure 14, which plots the cumulative number of nucleotide changes in seven proteins against the paleontological dates of divergence of 17 species of mammals. The overall rate of nucleotide substitution is fairly uniform. Some primates (shown by the dots at the lower left of the Figure) appear to have evolved at a slower rate than the average for the rest of the species. This anomaly occurs because the more recent the divergence of any two species the more likely it is that the changes observed will depart from the average evolutionary rate. As the length of time increases, periods of rapid and slow evolution in any lineage are likely to cancel one another out.

The neutrality theory of molecular evolution. In the late 1960s it was proposed that at the molecular level most evolutionary changes are selectively "neutral," meaning that they are due to genetic drift rather than to natural selection. Nucleotide and amino-acid substitutions appear in a population by mutation. If alternative alleles (alternative DNA sequences) have identical fitness—if they are identically able to perform their function—changes in allelic frequency from generation to generation will occur only by genetic drift. Rates of allelic substitution will be stochastically constant; that is, they will occur with a constant probability for a given gene or protein. This constant rate is the mutation rate for neutral alleles.

Alternative alleles with identical fitness

According to the neutrality theory, a large proportion of all possible mutants at any gene locus are harmful to their carriers; these mutants are eliminated by natural selection, just as standard evolutionary theory postulates. The neutrality theory also agrees that morphological, behavioral, and ecological traits evolve under the control of natural selection. What is distinctive in the theory is the claim that at each gene locus there are several favourable mutants, equivalent to one another with respect to adaptation, so that they are not subject to natural selection among themselves. Which of these increases or decreases in frequency in one or another species is purely a matter of chance, the result of random genetic drift over time.

Neutral alleles are those that differ so little in fitness that their frequencies change by random drift rather than by natural selection. This definition is formally stated as $4N_e s < 1$, where N_e is the effective size of the population, and s is the selective coefficient that measures the difference in fitness between the alleles.

Assume that k is the rate of substitution of neutral alleles per unit time in the course of evolution. The time units can be years or generations. In a random-mating popula-

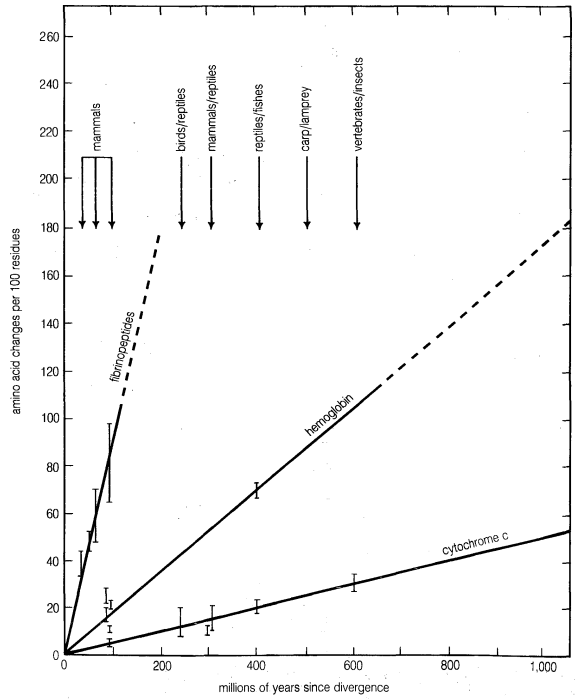


Figure 13: Three proteins with very different evolutionary rates: fibrinopeptides (very fast), hemoglobin (intermediate), cytochrome c (slow).

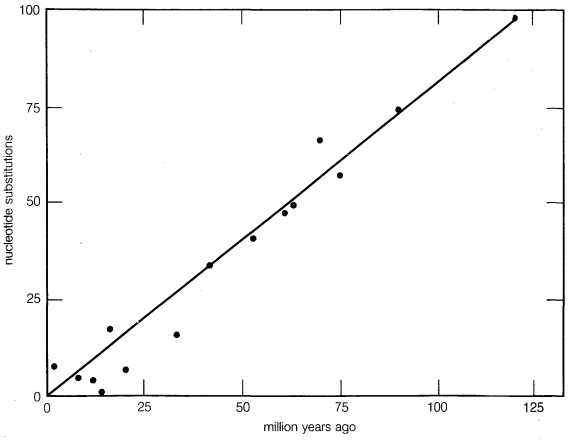


Figure 14: Rate of nucleotide substitution over paleontological time. Each dot marks (1) the point at which a pair of species diverged from a common ancestor and (2) the number of nucleotide substitutions, or protein changes, that have occurred since the divergence. The solid line drawn from the origin to the outermost dot gives the average rate of substitution.

From F.J. Ayala, E. McMullin (ed.), *Evolution and Creation* (1985)

tion with N diploid individuals, $k = 2Nux$, where u is the neutral mutation rate per gamete per unit time (time measured in the same units as for k) and x is the probability of ultimate fixation of a neutral mutant. The derivation of this equation is straightforward: there are $2Nu$ mutants per time unit, each with a probability x of becoming fixed. In a population of N diploid individuals there are $2N$ genes at each locus, all of them, if they are neutral, with an identical probability, $x = 1/2N$, of becoming fixed. If this value of x is substituted in the equation above ($k = 2Nux$), the result is $k = u$. In terms of the theory, then, the rate of substitution of neutral alleles is precisely the rate at which the neutral alleles arise by mutation, independently of the number of individuals in the population or of any other factors.

If the neutrality theory of molecular evolution is strictly correct, it will provide a theoretical foundation for the hypothesis of the molecular evolutionary clock, since the rate of neutral mutation would be expected to remain constant through evolutionary time and in different lineages. The number of amino-acid or nucleotide differences between species would, therefore, simply reflect the time elapsed since they shared the last common ancestor.

Debate over the neutrality theory

Evolutionists debate whether the neutrality theory is valid. Tests of the molecular clock hypothesis indicate that the variations in the rates of molecular evolution are substantially larger than would be expected according to the neutrality theory. Other tests have revealed substantial discrepancies between the amount of genetic polymorphism found in populations of a given species and the amount predicted by the theory. But defendants of the theory argue that these discrepancies can be assimilated by modifying the theory somewhat; by assuming, for example, that alleles are not strictly neutral, but their differences in selective value are quite small. Be that as it may, the neutrality theory provides a "null hypothesis," or point of departure, for measuring molecular evolution.

BIBLIOGRAPHY. Early seminal works of evolutionary theory include CHARLES DARWIN and ALFRED WALLACE, "On the Tendency of Species to Form Varieties, and on the Perpetuation of Varieties and Species by Natural Means of Selection," *Journal of the Proceedings of the Linnean Society*, 3(9):45–62 (1858); and CHARLES DARWIN, *On the Origin of the Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (1859), also available in many modern editions, and *The Descent of Man, and Selection in Relation to Sex*, 2 vol. (1871, reprinted in 1 vol., 1981). G. MENDEL, *Experiments in Plant Hybridisation* (1965; originally published in German, 1866), provides the groundwork for all subsequent studies in heredity, including R.A. FISHER, *The Genetical Theory of Natural Selection*, 2nd rev. ed. (1958); and J.B.S. HALDANE, *The Causes of Evolution* (1932, reissued 1966). THEODOSIUS DOBZHANSKY, *Genetics and the Origin of Species* (1937, reprinted 1982), is the classic foundation of the synthetic theory of evolution; see also JULIAN HUXLEY, *Evolution: The Modern Synthesis*, 3rd ed. (1974).

The history of evolutionary theories from Darwin to the present is traced in RONALD W. CLARK, *The Survival of Charles Darwin: A Biography of a Man and an Idea* (1984, reissued 1986), which also presents an engaging biography of Darwin. The most authoritative historical treatise of evolutionary ideas from antiquity to the present is ERNST MAYR, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (1982). ERNST MAYR and WILLIAM B. PROVINE (eds.), *The Evolutionary Synthesis: Perspectives on the Unification of Biology* (1980), contains historical articles by several of the great evolutionists who formulated the synthetic theory.

Modern treatments of evolutionary theory include G. LEDYARD STEBBINS, *Darwin to DNA, Molecules to Humanity* (1982), a readable discussion providing coverage of human evolution, both biological and cultural. A fairly comprehensive text requiring only general biology as background is FRANCISCO J. AYALA and JAMES W. VALENTINE, *Evolving: The Theory and Processes of Organic Evolution* (1979). A more advanced text is THEODOSIUS DOBZHANSKY *et al.*, *Evolution* (1977). FRANCISCO J. AYALA, *Population and Evolutionary Genetics: A Primer* (1982), provides an introduction to the genetics of the evolutionary process. A more advanced and mathematically demanding work is PHILIP W. HEDRICK, *Genetics of Populations* (1983, reissued 1985). The origin of species is the subject of MICHAEL J.D. WHITE, *Modes of Speciation* (1978); and of the more comprehensive ERNST MAYR, *Animal Species and Evolution* (1963), which is a classic work. G. LEDYARD STEBBINS, *Flowering Plants: Evolution Above the Species Level* (1974), discusses plant speciation and evolution.

A good introduction to the fossil record is a collection of articles from *Scientific American*, edited by LÉO F. LAPORTE, *The Fossil Record and Evolution* (1982). GEORGE GAYLORD SIMPSON, *The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man*, 2nd rev. ed. (1967, reissued 1971), is written for the general reader yet is an authoritative work dealing particularly with paleontological principles and the evolutionary process through time; somewhat more technical is his *Major Features of Evolution* (1953, reprinted 1969). An authoritative treatise on paleontological principles is STEPHEN JAY GOULD, *Ontogeny and Phylogeny* (1977).

Two good introductions to molecular evolution are FRANCISCO J. AYALA (ed.), *Molecular Evolution* (1976); and MASATOSHI NEI and RICHARD K. KOEHN (eds.), *Evolution of Genes and Proteins* (1983). The neutrality theory is presented in full by its main theorizer in MOTOO KIMURA, *The Neutral Theory of Molecular Evolution* (1983); and the theory that evolutionary changes happen not gradually but abruptly is advanced by one of its originators in NILES ELDREDGE, *Time Frames: The Rethinking of Darwinian Evolution and the Theory of Punctuated Equilibria* (1985).

(F.J.A.)